DOCUMENT RESUME

ED 478 789 PS 031 430

AUTHOR	Puma, Michael; Bell, Stephen; Shapiro, Gary; Broene, Pam; Cook, Ronna; Friedman, Janet; Heid, Camilla
TITLE	Building Futures: The Head Start Impact Study. Research Design Plan.
INSTITUTION	Administration for Children, Youth, and Families (DHHS), Washington, DC. Head Start Bureau.
PUB DATE	2001-03-31
NOTE	105p.
AVAILABLE FROM	Head Start Information and Publication Center, Order Fulfillment Department, P.O. Box 26417, Alexandria, VA 22313-0417. Tel: 703-683-2878; Fax: 703-683-5769; e-mail: puborder@headstartinfo.org; Web site: http://www.headstartinfo.org. For full text: http://www.acf.hhs.gov/programs/core/ongoing_research/hs/research_plan_4-251.pdf.
PUB TYPE	Reports - Evaluative (142) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC05 Plus Postage.
DESCRIPTORS	Compensatory Education; *Data Analysis; *Data Collection; Delivery Systems; Early Intervention; Experiments; Field Tests; Low Income Groups; Measurement Techniques; Measures (Individuals); *Outcomes of Education; *Preschool Education; *Program Effectiveness; Program Evaluation; *Research Design; Sampling; School Readiness
IDENTIFIERS	*Project Head Start; Reauthorization Legislation

ABSTRACT

Along with the rapid expansion over the past decade of Head Start, a program providing comprehensive early childhood development services to low-income children, their families, and their communities, has come the demand for rigorous research to demonstrate program effectiveness. This report describes the proposed design of a national study of the impact of Head Start on children, mandated through the 1998 reauthorization of Head Start. The study will entail the selection of a sample of Head Start applicants who will be randomly assigned to either a treatment or a control group. A field test is proposed to run parallel with initial site recruitment to inform later design modifications or refinements. The focus of the study will be the impact of federal Head Start on children, how program effects vary along a variety of dimensions, and the contribution of the Head Start service model to child outcomes compared to those of less intensive services. Other information included in the report are details of the research goals and objectives; the selection of study samples of grantees/delegate agencies and children; participant recruitment procedures; random assignment procedures; definition and collection of outcome measures related to children, families, programs, and context; and data analysis plans regarding treatment impact, non-response bias, and community-wide effects. The report's four appendices describe the sample cluster stratification, describe ways to analyze program impacts without placing a 2-year exclusion on controls, discuss data sources for Head Start programs in "saturation" communities, and contain the impact research-related amendment to the Head Start Act 1998. (KB)

**Reproductions supplied by EDRS are the best that can be made
from the original document.**

ED 478 789

Building Futures
Head Start
Impact Study

Building Futures: The Head Start Impact Study

Research Design Plan

Michael Puma
Stephen Bell
Gary Shapiro
Pam Broene
Ronna Cook
Janet Friedman
Camilla Heid

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

March 31, 2001

PS 031430

BEST COPY AVAILABLE

Table of Contents

TABLE OF CONTENTS .... **I**

1. RESEARCH DESIGN OVERVIEW **1**

1.1 STUDY BACKGROUND 1
Research Goals and Objectives *1*
1.2 OVERALL RESEARCH DESIGN 2
1.3 HOW DO WE DEFINE THE TREATMENT? 4

2. SAMPLING PLAN **8**

2.1 STEPS IN SAMPLE SELECTION 8
Step 1: Include all Head Start Grantees/Delegate Agencies and Children *8*
Step 2: Create Geographic Grantee/Delegate Agency Clusters (GGCs) *10*
Step 3: Stratify the Sample to Ensure National Program Representation *12*
Step 4: Select Sample of Geographic Grantee Clusters (GGCs) *20*
Step 5: Identify Grantees/Delegate Agencies Eligible For The Study *21*
Step 6: Select 75 Grantees/Delegate Agencies *23*
Step 7: Recruit Sites For The Study *26*
Step 8: Select 225 Head Start Centers From The Sampled Grantees/Delegate Agencies *27*
Step 9: Divide the Population of Head Start Children into One-Year and Two-year Participants *30*
Step 10: Select Appropriately-Sized Samples of Head Start Children *31*
2.2 SAMPLING ISSUES 33
Sample Attrition *33*
Unequal Allocation to Treatment and Control Groups *33*
Excluding Very High-Risk Children *34*
Placing 3-Year Olds Into The Control Group For Two Years *34*
Families (or Households) With Multiple Eligible Children *35*
Grantees/Delegate Agencies Do Not All Serve 3-Year Olds *36*
2.3 STATISTICAL POWER 36
Why Do We Need Such A Large Sample? *41*
Will The Sample Remain Representative Over Time? *42*

3. FIELD TEST PLAN **43**

3.1 INTRODUCTION 43
Purpose Of the Field test *43*
3.2. FIELD TEST DESIGN 45

4. SITE RECRUITMENT STRATEGY **50**

5. RANDOM ASSIGNMENT OF CHILDREN **54**

Informed Consent *58*
Monitoring Random Assignment *58*
Use of Incentives *60*

6. DATA COLLECTION **61**

6.1 OVERVIEW 61
6.2. DATA COLLECTION STRATEGIES 64
Planned Data Collection Activities *64*
Data Collection Sources *64*
How Will We Encourage Participation? *66*
How Will We Maintain Contact With Families? *67*

7. MEASURES 69

7.1. OVERVIEW 69
7.2. PLANNED MEASURES 69
Child and Family Measures 69
Program-level Measures 71
Contextual Measures 72
Summary 74

8. ANALYSIS PLANS 75

8.1 BASIC IMPACT ESTIMATES 75
8.2. IMPACTS ON THE "TREATED" 76
8.3. CHECKING FOR "NON-RESPONSE" BIAS IN THE EXPERIMENTAL ESTIMATES DUE TO GRANTEE EXCLUSION 80
8.4. MEASURING COMMUNITY-WIDE EFFECTS 84

APPENDIX A: SAMPLE CLUSTER STRATIFICATION 86

APPENDIX B: WAYS TO ANALYZE IMPACTS WITHOUT PLACING A TWO-YEAR EXCLUSION ON CONTROLS 89

NOTATION AND POTENTIAL EXPERIMENTAL GROUPS 89
THE CHALLENGE AND A FIRST RESPONSE 90
SOME CAVEATS 92
A LOWER-BOUND STRATEGY 93
A COMPLEMENTARY UPPER BOUND 95
POTENTIAL PROBLEMS WITH THIS APPROACH 96

APPENDIX C: DATA SOURCE FOR HEAD START PROGRAMS IN "SATURATION" COMMUNITIES, FACES 2000 97

APPENDIX D: IMPACT RESEARCH-RELATED AMENDMENT TO THE HEAD START ACT, 1998, PL 105-285 99

1. Research Design Overview

1.1 *Study Background*

Head Start provides comprehensive early childhood development services to low-income children, their families, and the communities in which they reside. Over the last decade the program has experienced significant growth, particularly as greater attention has been paid to the need for early intervention in the lives of low-income children. In fact, the recent FY2001 budget agreement included an increase of $933 million for Head Start, for a total annual funding of $6.2 billion. Along with this growth have come initiatives calling for improved outcomes and accountability. Head Start is not, however, alone in this—in an era of fiscal constraints, all Federal agencies are being challenged to demonstrate program results, not simply report on program staffing and processes.

During this rapid expansion of Head Start, the U.S. General Accounting Office (GAO) released two reports underlining the lack of rigorous research on Head Start's effectiveness noting in the 1997 report that "...the body of research on current Head Start is insufficient to draw conclusions about the impact of the national program."[1] The 1998 report added, "...the Federal government's significant financial investment in the Head Start program, including plans to increase the number of children served and enhance the quality of the program, warrants definitive research studies, even though they may be costly."[2]

Based upon the GAO recommendation, and the testimony of research methodologists and early childhood experts, Congress mandated through the 1998 reauthorization of Head Start that the Department of Health and Human Services (DHHS) determine, on a national level, the impact of Head Start on the children it serves. In October 2000, DHHS awarded a contract to Westat, Inc. in collaboration with The Urban Institute, the American Institutes for Research, and Decision Information Resources to conduct this research study.

Research Goals and Objectives

According to a report by the Advisory Committee on Head Start Research and Evaluation,[3] the Head Start Impact Study is intended to answer two overarching research goals or objectives :

- *"What difference does Head Start make to key outcomes of development and learning (and in particular, the multiple domains of school readiness) for low-income children?"*

---

[1] U.S. General Accounting Office (1997). ***Head Start: Research Provides Little Information on Impact of Current Program.*** Washington DC: Author.

[2] U.S. General Accounting Office (1998). ***Head Start: Challenges in Monitoring Program Quality and Demonstrating Results.*** Washington DC: Author.

[3] Advisory Committee on Head Start Research and Evaluation (1999). ***Evaluating Head Start: A Recommended Framework for Studying the Impact of the Head Start Program.*** Washington, DC: US Department of Health and Human Services.

- *"Under what circumstances does Head Start achieve the greatest impact? What works for which children? What Head Start services are most related to impact?"*

The first study goal can be broken down into the following two research questions:

1. What impact does Head Start have on children's: physical well-being and motor development; social and emotional development; approaches to learning; language development and emerging literacy; and, cognition and general knowledge?

2. What impact does Head Start have on parental practices that contribute to children's school readiness?

The second goal of the study will involve an examination of various factors that may be related to higher (or lower) program impacts, i.e.: How do program impacts vary among….

1. different types of children, e.g., gender, race/ethnicity, age cohort (3- vs. 4-year olds), presence of disabilities?

2. children from different home environments, e.g., family composition, income, parental practices related to school readiness?

3. grantees/delegate agencies with different characteristics, e.g., overall program "quality," part- vs. full-day services, grantee auspice, a 1- vs. 2-year program of services, group size and child-adult ratio, staff characteristics, teacher-child interactions, and the program's "instructional focus?"

4. different types of childcare and preschool environments, e.g., the availability and quality of alternative care settings, and state and local government resources for, and regulation of, childcare and "Head-Start-like" preschool education programs?

1.2 Overall Research Design

As noted above, the primary purpose of this study is to determine whether Head Start has an "impact" on participating children, and if so, if impacts vary as a function of the characteristics of children, their families, Head Start grantees/delegate agencies, and their environments. By impact we mean the difference between outcomes observed for Head Start participants and what *would have been observed for these same individuals had they* ***not*** *participated in Head Start*.

The key question, then, is how do we determine what outcomes would have been observed if the children had not participated in Head Start? In many studies, researchers have used a variety of methods to construct a "participant-like" comparison group, but even the best attempts at this have significant drawbacks primarily related to what evaluators call "selection bias," i.e., the extent to which program participants are determined, or selected, by a process that makes them different from non-participants on factors that are often difficult to measure but that lead to different outcomes independently of the Head Start "treatment."

To avoid such difficulties, Congress and the Advisory Committee on Head Start Research and Evaluation (1999) recommended the use of a **randomized research design** for the Head Start Impact Study. As such, the study will involve the selection of a sample of Head Start applicants who will then be randomly assigned to either a ***treatment group*** (these children and their families will receive Head Start services) or to a ***control group*** (these children will receive other available services selected by their parents). Under this randomized design, a simple comparison of outcomes for the two groups yields an unbiased estimate of the impact of the treatment condition—for example, the effect of Head Start on children's social and emotional school readiness.

The advantage of this research design is that if random assignment is not severely compromised by either the individuals in the study (e.g., high rates of "no shows," treatment group members who do not enter the program, or "cross-overs," control group members who manage to obtain Head Start services), or by grantee/delegate agency staff (e.g., using other criteria rather than random assignment to decide which individuals do and do not receive services), program participants should not differ in any systematic or unmeasured way from non-participants. More precisely, there will be differences between the two groups, but the expected or average value of these differences is zero (i.e., selection bias is removed by random assignment).

Within this framework of a randomized study design, the project must meet three requirements:

- The study must produce ***internally valid*** estimates of the impact of Head Start on the children of low-income-eligible families relative to what they would have received in the absence of Head Start. The counterfactual will, however, involve a comparison to alternative types of services available in their respective communities rather than to "no services," since parents of children assigned to the control group will be free to choose their own locally-available care arrangements.

- The study should strive for high ***external validity***. That is, the resulting impact estimates should, to the extent possible, represent the national average impact of Head Start on children served.

- The study should also include an assessment of the ***variation in impact estimates*** for different types of children, and for different types of grantee/delegate agency characteristics and contextual circumstances.

These three requirements impose competing demands on the study design task. For example, a design that provides strong internally-valid estimates of program impact may not be the optimal design for examining how those impact estimates vary across a variety of grantee/delegate agency characteristics. As a consequence, the design task for this project demands that a number of trade-offs be made so that questions focusing on overall program impact can be reliably answered (i.e., does the program "work?"), while at the same time allowing for the ability to explore questions about where and for whom it "works" best.

In light of these concerns, the Advisory Committee recommended a field test to gain additional knowledge regarding any possible design option that would be proposed. The current study proposes such a field test beginning in the Spring of 2001, and running somewhat parallel with the initial recruitment of sites for the full study (in which full implementation including recruitment and random assignment of the main sample of families will not occur until the Summer of 2002). The proposed sampling plan for the full study is first described in Section 2 and the field test is described in Section 3. The timing of these early study activities will allow multiple opportunities to take advantage of knowledge gained from the field test (or from efforts to early recruit sites for the main study) to modify or further refine the design of the full-scale study, as necessary.

1.3 How Do We Define The Treatment?

Before moving on to a description of how we plan to meet these study requirements, we need to be clear about our definition of the Head Start "treatment." This is important because we will be randomly assigning some children (those in the treatment group) to get "it" and some to get anything else that is "not it" (those in the control group). Moreover, because our estimate of program impact will be the difference in outcomes between these two groups of children we need to be clear about what we are measuring (and, later, what we have learned from our analysis results).

The legislative mandate for this study requires "...a national analysis of the impact of Head Start programs" for federal policy making purposes. By inference, we take this to mean the impact of the **federal** Head Start programs as they **currently** operate—i.e., services offered by organizations receiving federal Head Start funds that are required to meet (and are monitored against) the regulatory requirements of the Head Start Performance Standards when serving Head Start-eligible children.

This focus on federal Head Start service providers, program models, and performance standards as currently configured is clearly reflected in the report of the Advisory Committee on Head Start Research and Evaluation (1999). On the one hand, the Committee describes the "program" as consisting of "Head Start grantees" (p.5) that "..must meet a set of Program Performance Standards that define the core services that Head Start programs are required to provide," and that "a monitoring and technical assistance effort ensures that programs are in compliance with the Performance Standards" (p.7). On the other hand, the Committee explicitly rejected a study design "which randomly assigned sites to Head Start as it is now or to Head Start enhancements" (p.57), clearly focusing the study on how Head Start currently operates, not as it could operate.

From this discussion we derive the following definition of the Head Start treatment:

- *A program of services, provided by a Head Start grantee/delegate agency to Head Start-eligible children, that is....*

    1. *Funded at least in part by federal Head Start dollars, and*

    2. *Required to adhere to—and monitored against—the full range of Head Start performance standards.*

Imbedded in this definition is the assumption that the Head Start "program" is defined first and foremost by the comprehensive **service package** it mandates and monitors, not just the portion of that package **financed by federal dollars**.

This definition has, however, important implications for the control—or non-Head Start—group that comprises the other half of the study population, i.e., anything that is not part of our definition of the treatment would be considered a non-Head Start experience and, hence, appropriate for control group members. This would include:

- Head-Start-like services that are **not** funded at least in part with federal Head Start dollars but which may be fully monitored against all Head Start Performance Standards (e.g., a small percentage of the slots funded by the State-funded Head Start program in Ohio).

- Head-Start-like services that are **not** funded with federal Head Start dollars and are **not** required to meet **all** Head Start Performance Standards.

- Non-Head-Start-like services including other preschool or childcare centers, family day care homes, and care provided in the parents home or by a relative.

It should be noted that some of the children supported by state Head Start funds will be located in federally-supported and monitored Head Start programs and, therefore, will be subsumed under the treatment definition proposed for this study. On the other hand, some control group members could receive a combination of these services, or no early childhood development services at all depending on what non-Head Start-funded services are available in their communities.[4] Indeed, they might not participate in any form of non-parental care, if that is the "natural" alternative to Head Start-funded services for a particular family in a particular setting. Families of control group children would simply make whatever other arrangements make sense for them among the available alternatives—which could include full-time parental care or any other day care and pre-school services within their communities.

---

[4] To illustrate this point, consider Ohio and Oregon, two states with "Head-Start-like" state funded programs. In Oregon, 3,024 children are in "state Head Start" but only 603 of these children served in 6 grantees receive ONLY state funds and there are no federally-funded Head Start programs in the same locations. Similarly, Ohio funds 22,066 children through its state Head Start initiate, but only 3,066 of these children in 3 grantees receive ONLY state funds, and again this occurs in communities without a federally-funded Head Start program.

As noted by the Advisory Committee, "In some Head Start communities the children who would comprise a control or comparison group are already in other care situations that reflect to varying degrees the Head Start Program Performance Standards. …thus the care provided in these settings may be very much like that provided through the Head Start program." But the picture is, in fact, even more complicated than envisioned by the Committee due to the growing move toward "blended" programs that further blur the distinctions between "treatment" and "control." A few examples will help make this clear:

- A Head Start grantee could use the federal funds to pay for a morning Head Start program and use state pre-K dollars to pay for an afternoon program, with no noticeable difference to the children or parents since the teachers and the program of services remain the same over the entire day.

- The same financial arrangement as above, but the teachers change between the morning and afternoon (the children remain in the same room), or the children move to another location with different teachers.

- Head Start federal funds are used to pay for part of the required comprehensive services, and other funds (e.g., subsidized childcare) are used to pay for other services that in total meet the Head Start Performance Standards. For example, childcare funds could support full-day classroom services, with Head Start funds used for support services (parent and family services, enhanced health services, consultants, etc.). Or, Head Start funds could pay for part of the services, and special education funding could be used to add a teacher (or teachers) to provide specialized services and to mentor the other teachers.

As a consequence, we need to modify our definition of the treatment as follows:

- *For the purposes of this study, to be considered part of the Head Start treatment, services must be provided solely to children who are administratively counted as "Head Start children" (and where such services are federally monitored against the program Performance Standards) or to combinations of Head Start and non-Head Start children where those children are commingled in the same classroom(s).*

This definition includes in the treatment condition those services funded by a mix of federal Head Start dollars and other program funds even if the federal share of support is quite small.

Our plan then is to define the treatment and control group as we have discussed above and to randomly assign children from the pool of applicants to one or the other alternative. The national impact estimate (i.e., the first goal of the study), derived from this experimental design, will then answer the question of the effect of the federal Head Start program on children. But, we do not plan to end our study at that point. Because we will have observed Head Start in its full range of contextual settings, under the second goal of this study we will be able to examine how program impacts vary along a variety of dimensions, including the extent to which there is wide availability of state Head Start and Head-Start-like programs in the community. Moreover, as discussed in our Analysis Plan, we will be able to use quasi-experimental techniques that capitalize on the randomized design to estimate the contribution of the Head Start service model to child outcomes compared to less intensive services even in places where control group members in some instances received non-federally funded Head Start-like services.

The remaining sections of this paper describe how we plan to select the study samples of grantees/delegate agencies and children, recruit study participants, conduct random assignment, define and collect our outcome measures, and conduct the impact analysis.

2. Sampling Plan

2.1 Steps in Sample Selection

As outlined in the legislative mandate, the Head Start Impact Study must provide "…a national analysis of the impact of Head Start" based on the selection of Head Start grantees/delegate agencies that "…operate in the 50 states, the Commonwealth of Puerto Rico, or the District of Columbia and that do not specifically target special populations." Furthermore, the Advisory Committee recommended that the sample of Head Start grantees/delegate agencies should reflect variation in a variety of characteristics including, "…region of the country, race/ethnicity/language status, urban/rural, and depth of poverty in communities," and "…design of program as a one-year or two-year experience for children; program options (e.g., center-based, home-based, part-day, full-day); auspice (e.g., Community Action Agency, public school, non-profit organization); community-level resources; alternative childcare options for low-income children; and, the nature of the childcare market and the labor market in the community studied."

In this section, a 10-step plan for selecting the study sample is described to meet these joint requirements of national impact estimates and an assessment of the variation in program impacts along a variety of child/family, program, and community dimensions. Exhibit 1 provides a summary of the planned sampling approach

Step 1: Include all Head Start Grantees/Delegate Agencies and Children

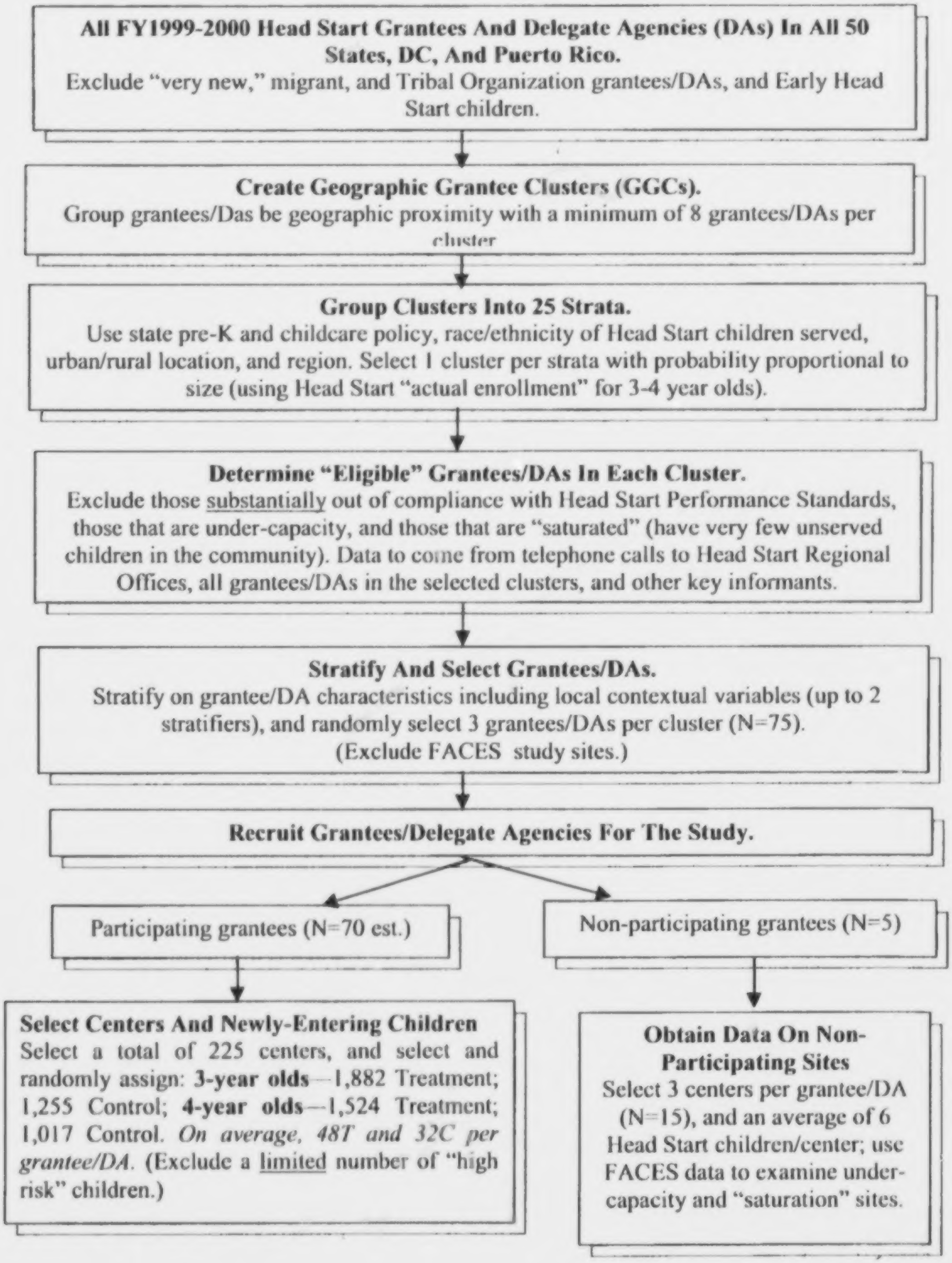
While the Advisory Committee discussed several options for selecting sites, the current study design, a stratified national sample (with limited program replacement), was determined to be the strongest of the recommended options. The sampling strategy begins with the legislative-mandated requirement, under the first study goal, to have a national impact estimate that captures the wide variety of grantees/delegate agencies operating in all 50 states, the District of Columbia, and the Commonwealth of Puerto Rico, and that do not specifically target special populations. In consideration of the final requirement, several initial exclusions have been incorporated into the design that can be defined at the outset of the sampling process[5]:

- grantees/delegate agencies specifically serving migrant children;
- Head Start programs operated by Tribal Organizations;
- children enrolled in Early Head Start (i.e., those younger then 3 years of age); and,
- as recommended in the Advisory Committee report (1999), grantees/delegate agencies that are "extremely new to the program" because they may not represent stable Head Start operations due to normal early startup problems.

---

[5] Several other exclusions are discussed below which cannot be easily defined in advance for the universe of Head Start grantees/delegate agencies and children.

Exhibit 1: Overall Plan For Sample Selection



All FY1999-2000 Head Start Grantees And Delegate Agencies (DAs) In All 50 States, DC, And Puerto Rico.
Exclude "very new," migrant, and Tribal Organization grantees/DAs, and Early Head Start children.

Create Geographic Grantee Clusters (GGCs).
Group grantees/Das be geographic proximity with a minimum of 8 grantees/DAs per cluster

Group Clusters Into 25 Strata.
Use state pre-K and childcare policy, race/ethnicity of Head Start children served, urban/rural location, and region. Select 1 cluster per strata with probability proportional to size (using Head Start "actual enrollment" for 3-4 year olds).

Determine "Eligible" Grantees/DAs In Each Cluster.
Exclude those substantially out of compliance with Head Start Performance Standards, those that are under-capacity, and those that are "saturated" (have very few unserved children in the community). Data to come from telephone calls to Head Start Regional Offices, all grantees/DAs in the selected clusters, and other key informants.

Stratify And Select Grantees/DAs.
Stratify on grantee/DA characteristics including local contextual variables (up to 2 stratifiers), and randomly select 3 grantees/DAs per cluster (N=75).
(Exclude FACES study sites.)

Recruit Grantees/Delegate Agencies For The Study.

Participating grantees (N=70 est.)

Non-participating grantees (N=5)

Select Centers And Newly-Entering Children
Select a total of 225 centers, and select and randomly assign: **3-year olds**—1,882 Treatment; 1,255 Control; **4-year olds**—1,524 Treatment; 1,017 Control. *On average, 48T and 32C per grantee/DA.* (Exclude a limited number of "high risk" children.)

Obtain Data On Non-Participating Sites
Select 3 centers per grantee/DA (N=15), and an average of 6 Head Start children/center; use FACES data to examine under-capacity and "saturation" sites.

The starting point for creating this initial population of Head Start grantees/delegate agencies[6] will be the 1999-2000 Program Information Report (PIR) database maintained by ACYF. Migrant and Tribal Organization grantees/delegate agencies can be readily identified from this database, and "new" programs will be identified as those grantees/delegate agencies that were listed in the 1999-2000 PIR but which were **not** listed in the 1998-1999 PIR (i.e., eliminating grantees/delegate agencies that were in operation for approximately less than two years)[7]. Early Head Start children will be identified and excluded once the sample of grantees/delegate agencies has been selected (see below).

Step 2: Create Geographic Grantee/Delegate Agency Clusters (GGCs)

Once the initial list of Head Start grantees/delegate agencies has been assembled, we will cluster grantees/delegate agencies based on their geographic proximity, and subsequently select a sample of geographic grantee clusters (GGCs). We have elected to initially select clusters of grantees/delegate agencies—rather than selecting a simple random sample from the PIR list— to reconcile two competing needs:

- ***The Need For a Large Sample of Grantees/Delegate Agencies:*** One of the major research objectives involves the determination of "what works best for whom." This requires larger samples of **grantees/delegate agencies** than would otherwise be needed to answer the first major research question—"what is the national impact of the Head Start program." That is, knowing what works best requires having a sufficient number of diverse grantees/delegate agencies, not just simply large samples of children, covering the range of dimensions along which Head Start services may vary and that have consequences for participating children.

- ***Logistical Constraints:*** The random selection of more grantees/delegate agencies means collecting data in more communities and this can both increase costs and potentially decrease the quality of the data. That is, maintaining the integrity of random assignment, and ensuring the collection of high quality data, requires close "hands on" and frequent interaction with local Head Start staff. To address this need, the plan is to assign a local site coordinator to manage the ongoing study activities in each of the selected clusters. If we were to disperse the sites more broadly, the cost of hiring and training the requisite number of coordinators would be prohibitive. The potential benefits of dispersing sites do not, in our view, outweigh the effective and efficient approach of coordinating a data collection effort that requires a hands-on approach.

The first step in this process will, therefore, use the PIR data to determine, for each county in the US, the number of grantees/delegate agencies with business addresses in the county and the actual number of 3-4 year old Head Start children they serve.[8] In some cases a grantee/delegate agency will have centers in more than one county. If such an agency is selected for the study sample, we will include in the sample all of the centers and children served by the grantee/delegate agency. This will increase travel costs somewhat, but will ensure full coverage of centers and grantees so that we have known probabilities of selection for all centers and children—this is important because we will eventually "weight up" our impact estimates to national averages.

---

[6] Grantees that provide direct services to children are considered to be a separate "program" from any delegate agencies they may operate; similarly, grantees that do not provide direct services are not included for sampling purposes.

[7] Theses identified new programs will be verified with the Head Start Bureau. Current estimates that these total about 53 grantees/delegate agencies. At a later point in the sample selection, when centers are selected from within grantees/delegate agencies, we will also exclude "extremely new" Head Start centers for the same reasons.

Each county in the United States that contains one or more grantees (and, as a consequence, each Head Start program and participating child) will be included in one of the created grantee clusters. However, we have set a minimum size of eight (8) grantees/delegate agencies per cluster to ensure that we have a sufficient number to meet our plan of selecting an average sample of three (3) grantees/delegate agencies per cluster (see below).[9]

Very small grantees/delegate agencies pose a particular difficulty, in that it will be difficult, if not impossible, to reach the desired sample size of centers or children for the smallest cases. Consequently, we will combine each grantee/delegate agency with fewer than 90 total 3- and 4-year old Head Start children (as reported in the PIR database) with another grantee/delegate agency in the same county if possible or in an adjacent county. Small grantees/delegate agencies will be combined with either another small grantee/delegate agency or with a "large" grantee/delegate agency. Such combinations can also involve multiple "small" grantees/delegate agencies if needed to meet the minimum criterion of at least 90 children.[10]

The actual cluster formation will be done using a proprietary Westat computer algorithm that combines counties into clusters so as to minimize the largest possible distance between any two points within the GGC. For example, the program first looks at combining county A with adjacent county B and using latitude/longitude information calculates the greatest distance between any two points in the two counties. This process is repeated for other adjacent counties that, together with county A, provide at least eight grantees/delegate agencies. Because many counties do not have a "resident" grantees/delegate agency (although children may be served by Head Start centers) the formation of clusters will, in some cases, require combinations of non-adjacent as well as adjacent counties. In these situations, clusters will be formed by hand. All cluster formations will be reviewed before proceeding to the next step in the sampling process.

---

[8] Addresses of individual Head Start centers are not available from a national data source for use in forming clusters; "actual enrollment" of 3-4 year olds from the PIR is used which includes children who have been enrolled for any length of time including drop-outs.

[9] In a few cases, where the geographic area for a GGC is very large (over 500 miles), we will permit a maximum of 7 grantees/delegate agencies in a cluster. Where possible, we have also taken into account physical boundaries (e.g., mountains, rivers) that would make data collection infeasible in the creation of clusters.

[10] There are only relatively few small grantee/delegate agencies. Ninety percent of them are estimated to have 76 or more enrollees, and seventy-five percent have more than 154 enrollees. The median size is 291, while the mean size is 458.

This plan will ensure that each Head Start grantee/delegate agency and participating child has a known probability of selection into the study sample, and—at later points in the sampling process—that the probabilities will be approximately the same for each child.

Step 3: Stratify the Sample to Ensure National Program Representation

The next step in the process will be to combine the GGCs into a total of 25 strata, each stratum having approximately the same number of 3- and 4-year old Head Start children across the clusters it contains. The use of equal-sized strata will ensure that all Head Start children in the nation have an equal chance of inclusion in the sample when, at a later step in the process, a single GGC is selected—with probability proportionate to size—to represent each stratum.

There are at least three reasons for stratifying clusters (and later grantees/delegate agencies) before randomly selecting those that will be included in the Head Start Impact Study:

- ***Statistical Estimation*** — Stratification, by reducing the variance within strata, increases the precision of overall population estimates compared to what would result from a simple random sample. Stratification combines observations that are similar and, as a consequence, each stratum can be represented by fewer observations (alternatively, holding sample size constant, each stratum can be represented with greater precision). If it were possible to make a completely homogeneous stratum (all members identical with respect to important characteristics), a single observation would serve as well as a large sample in representing the stratum.

- ***Population Representation*** — A secondary purpose of stratification, which is related to the precision of estimation, is to ensure that a particular sample is not far off from the overall population distribution. In other words, stratification helps ensure that the mix of different types of Head Start programs and children in the study sample comes close to matching the true mix in the total universe of Head Start programs. A well-stratified Head Start sample will have, for example, large and small grantees/delegate agencies, Hispanic and African American children, and different types of pre-kindergarten settings all represented in the sample roughly proportional to their distribution in the Head Start population.

- ***Analysis*** — Finally, stratification helps ensure that the sample has sufficient numbers of observations for analytically important subgroups (i.e., representation of the analytical domains of interest). For example, if we want to examine how impacts vary across different types of communities, we will need to ensure that we have adequate sub-samples to allow this type of comparison. Similarly, if we want compare program impacts for African American and Hispanic children, then the use of race/ethnicity as a stratifier ensures that we will not end up, due to bad luck, with a sample with, for example, very few Hispanic children.

In terms of statistical estimation and population representation, the best stratifiers are characteristics that are highly correlated with analytically important survey statistics. Often, the same factors define the domains of greatest analytic interest. For example, if we had measures of impacts on academic test scores, or on scales of social development, for each grantee/delegate agency (and thus each grantee cluster), these would be the perfect stratification variables for an impact study.

To obtain the desired increases in statistical precision, strata should be as homogeneous as possible. That is, stratification should create groups that are similar on characteristics that are of analytical interest; conversely, groups should be heterogeneous across strata. To gain "representativeness," strata should be designed to (1) include the range of differences that exist among sampling units, and (2) include them roughly in proportion to their distribution in the population — not too few nor too many. Of course, in situations where one wants to ensure inclusion of "rare" observations, stratification can also be used to "over-represent" them in the sample. Finally, where it does not conflict with the first two criteria, stratification should occur along dimensions that are important for analytical purposes. That is, strata should be designed to ensure the inclusion of subgroups for which separate analytical results are desired (e.g., by the race/ethnicity of Head Start children).

Operationally, stratification has to be based on information that is known for ***all*** units in the sampling frame; i.e., we must be able to classify all of the sampling units into the groups defined by the selected strata.[11] In a national study, this means confining stratifiers to factors for which national data are available, and for which national data can be disaggregated to the level of the chosen sampling units (e.g., grantees/delegate agencies).

Choosing Cluster Stratification Variables

There are two levels at which we plan to do stratification (and subsequent sample selection) for the Head Start Impact Study: (1) for geographic clusters of grantees/delegate agencies, and (2) for grantees/delegate agencies within selected clusters. The latter step (described below) will allow the incorporation of information that is unavailable nationally, but that can be collected once attention has been narrowed to the 25 sampled clusters.

As discussed above, stratification variables should be (1) related to expected variation in program impact, (2) representative of important analytical domains, **and** (3) capable of being measured for all clusters. One category of such variables is associated with the geographic location in which the grantees/delegate agencies operate:

- ***Region of the country*** — there is a wealth of data showing differences across regions on a host of characteristics that may affect program impacts including the historical pattern of how Head Start has developed and spread nationwide, child and family population characteristics, and economic and social trends (including sudden shocks such as economic downturns and immigration).

---

[11] Some misclassification is inevitable and can be handled during the post-survey weighting process.

- ***Urban location*** — important differences also exist between urban, suburban, and rural communities in terms of population characteristics, economic conditions, the determinants of poverty, and the availability of services for low-income children. Distinctions between the central cities of large metropolitan areas and other urban neighborhoods may also be important.

- ***State policy context*** — there are at least two aspects of state policy context that are likely to matter for this study:[12]

    1. **Comprehensive services for low-income children** — the extent to which states (or other non-Head Start sources) offer comprehensive preschool programs for low-income children that provide services that are "close to" those provided by Head Start. The availability of these services for non-Head-Start children is expected to lower the measured impact of the federal Head Start program. So the "quality" of services available to control group children, and the extent of their availability— factors again correlated with the expected magnitude of estimated program impacts and, as such, are important stratification variables.

    2. **Early elementary school** — similarly, the "quality" of elementary school experiences is likely to be highly correlated with the extent to which Head Start gains are sustained through the end of 1st grade. Higher expectations for educational outcomes on the part of state government (e.g., standards and accountability) may place greater demands on preschool education and thus raise the quality of available options.

- ***Need for services*** — the ability to produce program impacts may also vary with the level of need for comprehensive services in the population served.. For individual children—and thus for the Head Start service population generally—available data on indicators of need would include the economic status of the 's child's family and the child's own developmental status (compared to an age-appropriate norm). Since all children served by Head Start come from families at or below the poverty line, meaningful variation in economic need between one state's Head Start clientele and another's can only be measured with more detailed income data or other indicators of economic privation such as food insecurity.

Of these possible stratifiers, region of the country, urbanicity, and state policy regarding comprehensive pre-kindergarten programs are the best candidates for cluster stratification. State policy on K-12 education would seem to be more distal from our primary focus in this study (broad measures of "school readiness"). State and local indicators of service need—either economic or developmental—do not exist apart from the overall poverty rate. While important in thinking about how many Head Start "slots" might be funded, the aggregate poverty rate—or even the child poverty rate—provides little information about the factor of interest here: the *relative* need faced by the different collections of Head Start families and children actually served by the different programs.

---

[12] The same policy context factors matter at the local level, but local information on these factors is not available for all grantees/delegate agencies in the nation.

A second category of potential stratification variables includes factors that are associated with Head Start programs:

- ***Characteristics of Head Start children*** — the types of children being served is likely to have consequences for program impacts. Factors that may be important to consider include race/ethnicity, language(s) spoken in the home, parents' education and employment status, and parental involvement in the child's preparation for school.

- ***Characteristics of Head Start programs*** — although Head Start programs are all monitored against established Performance Standards, there are variations in program policies and actual operational characteristics that may affect program impacts, including: program auspice, full- vs. part-day program, the leadership of the center director, the educational background and skills of the classroom teacher, and the extent to which the program is focused on the full range of school readiness skills to be measured in the study.

Data on many of these child-level characteristics are unavailable at the cluster level. Information on the race/ethnicity of the children, however, does appear to be a good candidate for stratification.[13] Beyond aggregated data from the PIR, more detailed information on program-specific characteristics are generally not available for all grantees/delegate agencies nationally, and so are better suited for stratification within clusters once additional data can be obtained (see below).

Defining Cluster Stratification Variables

Based on the preceding discussion, four stratifiers will play a part in the selection of grantee clusters: (1) region, (2) urbanicity, (3) state policy regarding comprehensive pre-kindergarten programs, and (4) the racial/ethnic mix of the Head Start children served. The three geographically-based stratifiers will be based on the grantee/delegate agency business address, the only geographic information available for all grantees and delegate agencies nationally. The definition of each of the stratifiers is described below.

Region. Regions defined by the US Bureau of the Census are commonly used to capture geographic differences along a variety of dimensions, and were used in the design of the FACES study sponsored by ACYF. For the current study, however, we can do a better job of capturing geographic differences by using the 10 ACF regions as they are more likely to be associated with variations in program administration. The regional stratification will, therefore, be defined as follows:

---

[13] Prior research has shown that the size and longevity of children's gains through Head Start participation varies by race. For African-American children, gains may be quickly lost once children enter elementary school. See for example Currie, J. & D. Thomas, "Does Head Start Make a Difference?", ***American Economic Review***, June 1995.

- ***Northeast***: Head Start Regions 1, 2, and 3 — Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, New York, New Jersey, Puerto Rico, Pennsylvania, Delaware, Maryland, West Virginia, Virginia, and the District of Columbia, a total of 15 states.

- ***South***: Head Start Regions 4 and 6 — North Carolina, South Carolina, Georgia, Florida, Mississippi, Tennessee, Kentucky, Alabama, Louisiana, Oklahoma, Texas, New Mexico, and Arkansas, a total of 13 states.

- ***North Central***: Head Start Region 5 — Ohio, Indiana, Illinois, Michigan, Wisconsin, and Minnesota, a total of 6 states.

- ***Plains***: Head Start Regions 7 and 8 — Nebraska, Iowa, Missouri, Kansas, North Dakota, South Dakota, Montana, Wyoming, Colorado, and Utah, a total of 10 states.

- ***West***: Head Start Regions 9 and 10 — California, Arizona, Nevada, Idaho, Washington, Oregon, Alaska, and Hawaii, a total of 8 states.

The main differences vis-à-vis the Census regions are: (1) moving the mid-Atlantic states (DE, DC, MD, VA, and WV) from the Census South to our Northeast; (2) moving New Mexico from the Census West to our South; and (3) creating a new "Plains" region from portions of the Census West (CO, WY, UT, MT) and the Census North Central (NE, IA, ND, SD. MO, KS). **Clusters that cross regional boundaries will be assigned to strata on the basis of the category in which the largest percentage of the Head Start children in the cluster reside**, using the county of each grantee/delegate agency's business office as a proxy for the residential location of the children served. For example, if 60 percent of Head Start children in a particular cluster fall into the Northeast region and 40 percent into the South region, the cluster will be placed in the Northeast stratum

Urban Location. Stratification for urban location will be defined using the government designations for Metropolitan Statistical Areas (MSA's) and Beale Codes.[14] Three categories of urbanicity will be defined using this variable:

- a county **containing a central city of an MSA** with 1 million or more persons;

- a county in an MSA not included in the first category (i.e., a suburban county or any county in a small MSA); and,

- all other areas of the country (i.e., areas not in an MSA—predominantly small towns and rural).

---

[14] Beale codes classify counties into 10 different categories based on size of MSA and proximity to an MSA, as determined by the Department of Agriculture.

A cluster comprised of counties from two or more of these categories will be assigned to the stratum that contains the largest percentage of the Head Start children served by its grantees/delegate agencies, with the agency's business office address again used as a proxy for county of residence.

State Comprehensive Programs for Low-Income Preschool Children. Three potential sources of current information that attempt to describe state policy and programs that target low-income preschool children have been identified: Schulman, K., H. Blank, & D. Ewen (1999), ***Seeds of Success: State Prekindergarten Initiatives, 1998-1999,*** Washington, DC, The Children's Defense Fund; Cauthen, N.K., J. Knitzer, & C.H. Ripple (2000), ***Map and Track: State Initiatives for Young Children and Families***, NY, National Center for Children in Poverty, Columbia University; and, an effort currently underway at the Frank Porter Graham Child Development Center, University of North Carolina (UNC), Chapel Hill. Of these sources, only the Schulman, Blank, & Ewen (1999) report provides information that allows identification of comprehensive state programs that ***are similar to Head Start***. This distinction is vital because, as noted above, alternative programs most like Head Start are likely to have the largest (downward) effect on the magnitude of estimated program impacts. The ***Map and Track*** report identifies states that have created their own pre-kindergarten programs but does not identify the extent to which these programs provide "Head-Start-like" comprehensive services. The work being done at UNC disaggregates state-funded programs by where they are housed (e.g. in an education agency) but again does not identify how similar the programs are to Head Start.

This stratification variable will, therefore, be based on the work by Schulman, Blank, & Ewen (1999) and will consist of three categories:

- States with comprehensive state-funded pre-kindergarten programs **that are similar to Head Start**, i.e., "Head Start Performance Standards are followed…" or that have "Comprehensive service requirements that are the same as or similar to those of Head Start …." (p. 97). This group includes 18 states[15]: Alaska, Connecticut, Delaware, the District of Columbia, Hawaii, Kansas, Maine, Massachusetts, Minnesota, New Jersey, New Mexico, New York, Ohio, Oklahoma, Oregon, Rhode Island, Washington, and Wisconsin.

- States with state-funded pre-kindergarten programs that have **some comprehensive program components**, i.e., states that "have requirements that address health and/or social services to some extent" (p. 97) or "stress comprehensive services as one of their primary objectives" and "encourage pre-kindergarten programs to coordinate with health care, social services, and other agencies" (p. 99). This category consists of

---

[15] New Hampshire's "funding is used to supplement all Head Start programs, not to support individual slots" (p. 196). Because any "slot" supported in part or in whole by federal Head Start dollars will be considered part of the treatment for the purposes of the evaluation, none of New Hampshire's funds are expected to be used in serving members of the control group. Consequently, we have elected to NOT include it in this category but instead classify it among the states with no comprehensive services available to controls.

11 states: Alabama, Arkansas, Colorado, Florida, Georgia Iowa, Kentucky, Nebraska, Tennessee, Vermont, and Virginia.

- States meeting neither of the previous two requirements. This group consists of the remaining 22 states and Puerto Rico.

As with geographic region and urban location, any clusters that cross state lines based on grantee/delegate agency's business addresses will be assigned to strata here based on the category into which the largest percentage of Head Start children fall.

Race/Ethnicity. The final cluster stratification variable will capture differences across clusters in the proportion of African American, and Hispanic children served using the following three categories:

- ***High concentration of Hispanic Head Start children*** — the percentage of Hispanic children served by the grantees/delegate agencies in the cluster is at, or above, 40 percent;

- ***High concentration of African American children (but not of Hispanic children)*** — the percentage of non-Hispanic African American children served by the grantees/delegate agencies in the cluster is at, or above, 40 percent and the percentage of Hispanic children below 40 percent; and,

- ***Other*** — all other clusters not included in the preceding categories.

This is a more refined and stronger classification than that used by the FACES study to ensure minority representation. Unlike FACES, it will ensure a more representative sample for Hispanics as a separate group, and a more representative sample for African Americans as a separate group. Because these categories are derived by pooling the data on children across all grantees and delegate agencies in a cluster, the possibility seen earlier—in connection with other stratifiers—of clusters crossing category boundaries does not arise with this stratifier.

Sequencing. The goal, as previously described, is to create 25 strata from which a sample of **25 geographic clusters** will be selected with **probabilities proportional to size** (i.e., clusters with larger numbers of Head Start children will have a higher probability of being selected into the sample). The **strata will be constructed to be of approximately equal size**—i.e., they will each include about ***the same number of Head Start children.*** In this way, a set of 25 clusters drawn one from each stratum—followed by the selection of an **equal number of Head Start children** per cluster—will correctly represent the national population on the different stratification characteristics.

To see why this is the case, suppose 20 percent of Head Start children were in states with comprehensive state-funded pre-kindergarten programs similar to Head Start. Using this variable as a stratifier separates out the clusters from those states into their own strata. The requirement that all strata be of approximately equal size (in terms of the number of

Head Start children served) implies that there are 5 such strata (equal to 20 percent of 25 total strata for the nation). Picking one cluster per stratum puts 20 percent of all clusters in those same states—again, 5 of 25 total clusters. With equal numbers of children selected from each cluster, this translates into 20 percent of the *children* in the sample, commensurate with their share of the overall population. Without stratification—i.e., if we sampled 25 clusters at random from all over the country—we might be unlucky and obtain far less, or far more, than 20 percent of the child sample from this subpopulation.

Obviously, if we were to use all of the stratification variables and categories listed above, we would end up with more than 25 possible strata.[16] To avoid this, we will employ sequential stratification, and will also combine strata (i.e., collapse cells) with too few clusters to represent 4 percent of the nation's Head Start children (the share required from each of 25 equal-size strata). Clusters will first be stratified on the basis of the **availability of non-Head-Start comprehensive services**. Then within these three strata we will, where possible, stratify on the **3 racial/ethnic stratifiers**.[17] Next, we will incorporate, to the extent possible, the **5 geographic regions** and, finally, where possible, incorporate urban location. The definition of urban location categories will vary, however, according to the specific distribution of clusters in the cells where urbanicity can be used.

This sequence of stratification reflects our assessment of the relative importance of the four variables in terms of their relationship to likely program impacts. We give highest priority to the availability of non-Head-Start comprehensive pre-kindergarten services as the only variable that directly contributes to the likely treatment-control *difference* in service environment, the factor random assignment was designed to vary and on which all impact estimates will depend. In reality, it matters little whether this variable or race/ethnicity comes first, since both of the first two variables are likely to be fully accommodated in the stratification (they require just 9 divisions, each containing just 4 percent or more of the national population). Race/ethnicity is seen as second most important of the available stratifiers because of some evidence of potential differential impacts of Head Start by this factor in previous research (see above). There is no existing research evidence that either geographic region or urbanicity influence program effectiveness. Stratification by geographic region also seems quite important, however, to ensure "face validity" of the sample and to incorporate any variation in Head Start operating procedures that may arise among the different ACF regional offices. The last sequenced variable, urban location, is similarly important for "face validity" and to capture program operational differences. But, because of the restriction of 25 strata, it will be used only to the extent that the preceding stratifiers do not adequately account for differences in this dimension.

Final Comment. This type of stratification is commonly used to increase the precision of the national impact estimates, and to ensure that the sample represents the nation and is diverse along the dimensions that define the strata. This stratification is **not** done to give grantees/delegate agencies in particular strata higher probabilities of selection than those in other strata, or to ensure reliable sample estimates for each stratum. The stratification is only for the purpose of controlling the diversity of the sample while ensuring that we end up with: (1) a national sample of Head Start grantees/delegate agencies; (2) a national sample of Head Start participants; and (3) a sample that has a sufficient number of observations in the key analytical domains (e.g., to allow analysis of the variation in program impact among communities with different levels of availability of Head-Start-like programs for low-income children).

---

[16] That is: (5 regions) x (3 urban/rural locations) x (3 state prekindergarten environments) x (3 race/ethnicity categories) = 135 cells.

[17] Some divisions into strata will be constrained by the share of the population allocated to a cell at the previous level of stratification.

Appendix A provides a summary of the cluster stratification plan and shows the number of 3- and 4-year old Head Start children in each stratification cell.

Step 4: Select Sample of Geographic Grantee Clusters (GGCs)

Once the strata are formed, we will select one GGC from each of the approximately 25 strata with probability proportional to the total Head Start enrollment of 3- and 4-year olds in the GGC.[18] In the unlikely event that we are unable to obtain at least two or three eligible and participating grantees/delegate agencies from a GGC, we will select a replacement GGC from the same stratum.

Our decision to sample a total of 25 clusters for the primary study sample is based on a tradeoff between two competing demands — while clustering reduces costs and can improve our ability to control both random assignment and data quality, clustering also increases the variance of the impact estimates (i.e., leading to larger confidence intervals around the estimated program impacts). Unfortunately, the information needed to make precise calculations of the optimal number of GGC's does not exist *a priori*. However, based on our previous experience with similar studies (e.g., FACES, ECLS-K) we have determined that the optimal number of sampled grantees/delegate agencies is about 50-80, and that the optimal number of sample clusters is in the range of 19 (with about 4 grantees sampled per cluster) to 50 (with 1 grantee sampled per cluster). We have selected 25 sample clusters, with 3 grantees/delegate agencies[19] (for a total of 75 grantees/delegate) as a reasonable estimate of the optimal sample size. In our judgment, a reduction in the number of sample GGCs much below 25 would lead to large increases in variance; alternatively, an increase in the number of sampled clusters much above 25 would likely result in only modest variance reductions and larger relative impacts on cost and the quality of data collection.

---

[18] Selecting GGCs proportional to size is done to give clusters within each strata with more Head Start children a higher chance of being selected because we want an approximately equal sample size of Head Start children from each selected GGC, as well as approximately equal probabilities of inclusion for all Head Start children in the nation. This can be best achieved if greater weight is given to larger clusters because they should represent a greater proportion of the overall average program impact than very small areas. To make up for this greater weight given to larger concentrations of Head Start children, the sampling rate for children will be set lower in larger clusters at a later step in the sampling. If larger areas were not given greater weight at the point of selecting clusters, the number of sampled children within each selected larger cluster would have to be very high.

[19] Throughout the remainder of this paper we use the term grantee/delegate agency but, in fact, because we will (as noted above) combine small grantee/delegate agencies, a single sample selection may include a group of two or more grantees/delegate agencies.

Step 5: Identify Grantees/Delegate Agencies Eligible For The Study

As described above, the 25 sample GGCs will consist of at least eight grantees/delegate agencies. However, some of these grantees/delegate agencies are not eligible to participate in the randomized study, as recommended in the Advisory Committee report (1999). These exclusions could not be implemented as part of "Step 1" because they require information that is not centrally available for all Head Start grantees/delegate agencies. Specifically, we will exclude:

- "…sites that are out of compliance with Head Start standards;" and
- "…sites where Head Start saturates the community (i.e., where there are not enough unserved children to permit random assignment of a sufficient number of children to an unserved control group)."

The first group of grantees/delegate agencies that will be excluded—those that are substantially out of compliance—will be those that have deficiencies that are so serious as to warrant closure. We are not excluding "low quality" programs as we want to include the full range of the program's current operations. Rather, we will only exclude those grantees/delegate agencies that do not represent even the minimally acceptable level of operation. These out-of-compliance grantees/delegate agencies will be identified through conversations with the respective Head Start Regional Offices using a standard protocol to be developed in cooperation with the Head Start program monitoring staff. At a minimum, the criteria will include any program on a quality improvement plan (QIP) or formally designated as "high risk" by the respective regional office.

With regard to the second group of ineligible grantees/delegate agencies, there are actually three categories of program capacity that are relevant to this study:

- grantees/delegate agencies that are operating **in a saturated environment**, i.e., those that both have most or all of their available slots filled and are serving all of the children in the community who are eligible for, and wish to attend, Head Start;
- grantees/delegate agencies that are operating approximately **at full capacity but not in a saturated environment**, i.e., those that are serving all of the children that they can serve within their current capacity, **but** where there are more eligible children in the community than the agencies can accommodate; and,
- grantees/delegate agencies that are operating **substantially under-capacity**, i.e., those that have more available slots than they have enrolled children.

Grantees/delegate agencies in the third category will be determined after thorough exploration of the unserved population in the community.

The Advisory Committee has recommended that saturated grantees/delegate agencies be excluded from the study because it would be unethical to deny services when the grantees/delegate agencies are capable of providing Head Start benefits to all eligible children who apply. The second group of "full capacity grantees/delegate agencies in non-saturated communities" will be considered eligible to participate in the study.[20] The final group should also be excluded, as noted in the Advisory Committee report (1999), because there are unlikely to be sufficient numbers of possible control group children to support an adequate experiment. However, before deciding to exclude grantees/delegate agencies on this basis, we will work with ACYF and local program staff to determine if efforts could be implemented to expand local recruitment and enrollment to both fill available slots **and** provide sufficient numbers of children to provide the desired control group.

From a sampling perspective, it is important to characterize **all** of the grantees/delegate agencies in the selected 25 GGCs along the two dimensions described above, and to obtain some basic descriptive information about the excluded (i.e., ineligible) grantees/delegate agencies. The latter information will be important later during the analysis as a way to demonstrate the external validity of the estimated program impacts, and—if face validity is not as apparent as it might be—to identify where the study findings might be vulnerable to later criticism.

To collect the required information, study staff will call the respective Head Start Regional Offices, as well as each grantee/delegate agency in the 25 sampled clusters, and go through an established protocol of questions with each designated respondent. In addition, study staff will contact other agencies in the selected clusters that may be able to shed some light on the issue of grantee/delegate agency capacity including State pre-K directors and local Resource and Referral Agencies. The combined information will be used to make the necessary determinations of grantee/delegate agency eligibility for inclusion in the study.

What Does Full Capacity Mean in Practice?

A grantee/delegate agency does not have to be much over capacity to be eligible for inclusion in the Head Start Impact Study. As will be discussed below, our proposed sampling of Head Start children from many different grantees/delegate agencies, and their disproportionate assignment to the treatment group, requires that grantees have an eligible unserved population from which a control group can be drawn that represents, on average, **only about 7 percent of their current total enrollment.**[21] This is a very modest requirement in the average site which, we believe, will eliminate very few at-capacity grantees/delegate agencies—just those that currently serve almost all of their potential eligibles, and as such, imply only a small extension to the below-capacity group excluded from the experiment.

---

[20] Although no reliable data currently exist to divide operating grantee/delegate agencies into these three categories, our analysis of PIR data indicates that out of the 1,914 grantee/delegate agencies included in the FY1999 PIR, about 11 percent are under-capacity (i.e., 20% or more unfilled slots), and that the remaining 89 percent are at or above capacity (i.e., above 95% of funded enrollment). This gives us confidence that the "under-capacity" and "saturation" exclusions will be relatively small, and that we will be focusing the study on the vast majority of Head Start grantees/delegate agencies. Special analyses of "saturation" grantees/delegate agencies will also be conducted on a non-experimental basis (see below).

[21] As previously discussed, setting a minimum standard of at least 90 children per grantee/delegate agency, and combining small grantees, also helps reduce the implications of our need for a modest control group.

We plan to assess the extent of non-representativeness in the experimental sample caused by this step in the selection process using data from the ongoing FACES study, as described in the Analysis Plan below.

Step 6: Select 75 Grantees/Delegate Agencies

Once we have completed Step 5, we will have identified a pool of grantees/delegate agencies within each of the 25 clusters that are eligible for inclusion in the study. This group of grantees/delegate agencies will be **representative of Head Start grantees/delegate agencies nationwide that meet the criteria that we have established**. The next step in the sampling process will involve the selection of an average of three grantees/delegate agencies from within each of the 25 GGCs for a total of 75 grantees/delegate agencies that will be subsequently recruited for the study.

Initial Composition Checks

Before implementing this process, however, it is important to realize that although the clusters will have been selected to provide diversity along a variety of dimensions that will be important for this study, this is not automatically the case for the **eligible** grantees and delegate agencies. For example, removal of the "saturation" grantees/delegate agencies may skew the distribution, leaving a set of programs that do not meet the diversity goals of the initial cluster stratification. We will, therefore, need to check at this point that a sufficient number of Head Start children remain in each of the categories defined by the original 25 strata, by race/ethnicity, urbanicity, and non-Head Start policy contexts. It is possible, for example, that Hispanic children in urban environments with few "Head Start-like" program alternatives will be served disproportionately by "saturated" grantees and delegate agencies and, hence, under represented in the pool of eligible children. This does not necessarily mean that the eventual sample of children—consisting of a subset of the children served by eligible grantees/delegate agencies (see below)—will contain too few cases of this sort, but it does increase the odds of such an outcome. To guard against this hazard, we will check whether the drop-off in the number of eligible children in any stratum creates a serious risk that the final sample of children will contain too few observations for that cell.

Two further factors will need to be checked at this point: the total number of Head Start children in each cluster, and the division of participating children between 3-year-olds and 4-year-olds across the entire sample. With regard to total numbers, it is possible that one or more of the selected GGCs will contain too few eligible grantees/delegate agencies to yield the desired sample size once "saturation" grantees/delegate agencies are removed. Or it may be that the total number of eligible children is adequate but the distribution by age is skewed (since not all grantees and delegate agencies provide both 1-year and 2-year programs, or do so in unbalanced proportions). As with the cluster stratifiers, checks of the adequacy of the pool of "served" children by age (in eligible grantees/delegate agencies) will be conducted in the aggregate, across all the eligible grantees/delegate agencies in the 25 clusters. As long as each subgroup is present in sufficient numbers **somewhere** in the pool, the overall diversity goals of the sample will be met.

All of these checks relate to the children who **could** be selected into the sample once grantees/delegate agencies are chosen, not the smaller number eventually selected. Our opportunity to address potential sample shortfalls exists only at this early, "pre-selection" stage if we are to maintain known probabilities of selection for all children. Once specific grantees and delegate agencies are chosen on a probabilistic basis, additions or adjustments to the sample would reflect responses to the particular sample traits encountered, a step that alters the overall probability of selection for a given child in an unknown way. Adjustments made **prior** to selection of grantees/delegate agencies can be accommodated by adjustments in the analysis weights. Several adjustment strategies can be considered for "customizing" grantee/delegate agency selection, including sampling more than 3 grantees/delegate agencies from a particular cluster or over-sampling certain types of grantees/delegate agencies based on the number or composition of the children served. The goal here is to reinstate the mix of children chosen when sampling clusters to the set of children likely to be selected when sampling eligible grantees/delegate agencies.

Within Cluster Stratification

Depending on the demands of the "reinstatement" process, as described above, we may also be able to stratify eligible grantees/delegates before they are sampled. But at most we will have the ability to incorporate two new stratification variables at this point because of the number of grantees/delegates being selected. One possibility is to attempt to further increase the geographical clustering of the sample within each GGC. For example, suppose a sampled GGC consisted of grantees/delegate agencies spread throughout the state of Montana. Subject to the availability of a sufficient number of eligible grantees/delegate agencies, a methodology could be set up with a good chance of selecting either three grantees in the eastern portion of the state or three grantees in the western portion of the state.

Where we are less concerned about the geographic dispersion of the sample, we will employ stratification[22] on other factors, selecting from a broad range of possible grantee/delegate agency characteristics:

- ***Local context for pre-kindergarten services for low-income children*** — within clusters we will have a greater ability to "map the community" in terms of the availability of alternatives to Head Start services for control group children, and the extent to which the alternatives are similar to the Head Start service model. Again, the nature of this "counterfactual" world for the control group will have a great deal to do with our ability to detect program impacts, and their magnitude.

- ***Characteristics of Head Start children*** — as noted in the discussion of cluster strata, the types of children being served is also likely to have consequences for program impacts, including the number or extent of participation of non-English speaking children, and parental involvement in the child's preparation for school. These we would hope to be able to collect from individual grantees and delegate agencies in the selected clusters.

---

[22] We will likely employ "implicit" rather than "explicit" stratification for this step. In implicit stratification, grantees/delegate agencies will be sorted by the selected stratification variables, and we will then select a systematic sample from the sorted list based on an "every Xth case" selection rule

- ***Characteristics of Head Start programs*** — similarly, program policies and operational characteristics are also likely to affect program impacts, including: program auspice, full- vs. part-day program, the leadership of the center director, the educational background and skills of the classroom teacher, and the extent to which the program is focused on the full range of school readiness skills that we plan to measure in the study. Most of these factors can only be measured at the grantee/delegate agency or center level.

We will have access to a rich set of information sources at this stage, but our ability to incorporate additional stratification variables will be limited when selecting an average of just three grantees/delegate agencies within each cluster. The decision then comes down to a careful selection of the 1-2 ***most important*** factors that can also be measured with reasonable reliability (different stratification variables can be used in different clusters — there is no requirement that sampling be done consistently across clusters, as long as known probabilities are used).

At this point we have elected to defer a final decision on within-cluster stratification variables until we have selected our clusters and collected information about the different communities that are potential candidates for inclusion in the study. This will ensure that we make the best informed decision regarding the selection of the final sample.

Within Cluster Selection

Once we have created appropriate within-cluster strata and sampling methods, 75 **grantees/delegate agencies will be selected accordingly, wherever possible using probabilities proportional to size** (where the measure of size is the "actual" number of enrolled 3- and 4-year olds from the PIR). Thus, in most instances, the probability of being selected will be proportional to the overall size of the grantee/delegate agency. This does not exclude small- and medium-size grantees/delegate agencies, but rather—once child sampling probabilities are set inversely proportional to grantee size—we will have equal probabilities of selection for each child and, at the same time, approximately equal sample sizes across grantees. (As noted above, grantees with very few children will be paired with another grantee and sampling will be done for the pair, so as to ensure an adequate sample size of children.)

A final step in the selection of grantees/delegate agencies will be to, at the request of ACYF, minimize the sample overlap with the Head Start Family and Child Experiences Survey (FACES). Because no grantee had more than a 0.25 probability being selected in the FACES sample, we can statistically exclude them from the experimental study sample.[23]

Step 7: Recruit Sites For The Study

Site recruitment, described in the section below, will be an intensive effort that is expected to result in relatively few refusals or exclusions at the grantee/delegate agency level. But this is, of course, an empirical question that will be a major part of a decision to be made at the end of the recruitment process, in conjunction with information derived from the pilot study, to decide whether it is worthwhile to continue with the study as is, or to consider further refinements. Neither we, nor anyone else, can say with any confidence what fraction of selected grantees/delegate agencies will agree and be able to participate in the study, although the pilot study is expected to not only reveal a number of unanticipated challenges, but also help to develop various strategies to overcome some or all of these such challenges.

Of course, even if we are unsuccessful in our efforts to recruit a nationally representative sample for the randomized research design, ACYF could still decide to proceed with the recruited sites to address the impact of Head Start in multiple settings (i.e., the "medical model" discussed by the Advisory Committee), as well as the research questions under the second study goal — assessing variation in program impacts — which also do not necessarily require a nationally representative sample. Even under this perspective, the question arises of "how many participating grantees are needed to make the study worthwhile?" In our view, and this is the subject of the eventual feasibility decision, a 70-percent "participation rate" among selected grantees/delegate agencies (i.e., inclusion of about 50 grantees/delegate agencies in the study out of a target of about 75) would be acceptable for continuing, barring any systematic pattern of non-participation (e.g., all of the large urban grantees/delegate agencies refuse or are otherwise unable to participate). Under such circumstances, we would expect such study results to encompass a large enough range of relevant grantees/delegate agencies to support national impact estimation, and would be adequate for the "what works for whom?" research questions.

Replacement Sites

We will replace grantees/delegate agencies that are unable or unwilling to participate, **only to maintain the sample sizes needed to support important sub-group analyses.** For example, because we want to assess the extent to which program impacts vary as a function of the availability of Head-Start-like services in the community, we need to ensure that we have a sample that can support this type of analysis. However, apart from the need to restore sample sizes for certain key subgroups, it is not worthwhile to expend resources replacing the kinds of grantees and delegate agencies we **can never** bring into the study (non-participating programs) with more sites of the type **already included** in the study (those programs among the original selections that are willing to do random assignment).

---

[23] To eliminate bias that would normally result from such an exclusion, a weight adjustment will be made to account for the probability that a grantee was selected for FACES. For example, if a grantee had a 0.2 probability of selection for FACES, was not selected for FACES and was selected for this study, then it would be given a weight adjustment of 1/0.8.

We do, however, believe it is important to acknowledge what is missing from the study—sites unable or unwilling to conduct random assignment—and look at the ramifications of non-participation directly. Thus, we plan to adopt a three-pronged strategy in response to potential sample loss through grantees' inability to conduct random assignment:

- implement procedures to maximize the participation rate in the originally-selected sample including those derived from the feasibility test (see the following discussion);
- use replacement where necessary to ensure that a sufficient number of Head Start grantees/delegate agencies and children are included in the sample to hit our targets for sample size and variety of program and participant types and settings; and,
- collect data and conduct extensive analyses on the "lost" grantees to gauge the potential degree of "non-participation bias" in the impact estimates generated from the grantees/delegate agencies where random assignment is implemented.

With regard to the last point, we plan to implement the proposed measurement battery with a sample of Head Start children attending centers under the direction of **up to five of the non-participating grantees/delegate agencies** (see the separate discussion of Data Collection).[24] [25] When combined with similar information on grantees/delegate agencies operating in saturation communities collected as part of the FACES 2000 study, these data will encompass enough children to make the assessment of pre/post changes in child outcomes in the non-random assignment sites as precise statistically as our national impact estimates from the random assignment study sites. There will, of course, be no control group children in any of the non-random-assignment sites. (How these data will be used to examine potential exclusion biases is discussed in the Analysis Plan section below.)

Step 8: Select 225 Head Start Centers From The Sampled Grantees/Delegate Agencies

As discussed below, our estimates of expected minimum detectable differences in effect (MDDIEs) between Head Start programs with different characteristics—needed to answer the "what works for whom?" question—suggests that a sample of 225 distinct Head Start centers is needed for the Head Start Impact Study. When examining variation in program impact, Head Start centers (and classrooms within those centers) rather than grantees/delegate agencies are the pertinent unit of analysis because this is the level at which most program-level variation affecting children is likely to occur.

---

[24] We will collect data in up to five non-participating grantees/delegate agencies, sampling an average of 3 centers per grantee/delegate agency (the same average per grantee as in the random assignment sample) and an average of 6 children per center.

[25] To the extent that the final study sample includes less than 70 grantees/delegate agencies, the average samples of Head Start children selected in the remaining sites will be increased to preserve the size of the overall study sample.

The 225 Head Start centers will be selected[26] regardless of the number of participating grantees/delegate agencies (i.e., if we have our planned number of about 75 grantees/delegate agencies, we will select an average of 3.0 centers per grantee/delegate agency; if the sample drops to the minimum of, say, 50 grantees we will select an average of 4.5 centers per grantee/delegate agency). As also discussed below, the estimated MDDIEs are very insensitive to the degree of geographic clustering in the sample, so increased clustering due to higher grantee non-participation will have a negligible effect on our ability to address questions about "what works for whom?"

Ideally, we would prefer to select our sample of Head Start children for the study from the recruitment lists of all newly-entering children (see the subsequent discussion) without regard to the centers in which they are located — eliminating this added clustering would, in fact, improve the precision of the national impact estimates. But, this loss of clustering is, on the other hand, likely to result in higher data collection costs, as staff could be required to collect data at a large number of local sites that may be widely disbursed geographically.

On average, grantees and delegate agencies operate about nine centers each. [27] To contain data collection costs, our plan is to select an average sample of just three Head Start centers per grantee/delegate agency for a total of about 225 centers nationally. Centers would be selected with probabilities proportional to size (i.e., larger centers would have a greater chance of being selected for the study), and the planned sample of 80 children per grantee/delegate agency would be distributed evenly across the selected centers. Centers that are "saturated" (i.e., it would not be possible to find unserved children for a control group) would not be considered for inclusion in the study sample.

An important issue that arises regarding this plan, however, is the decision to spread the sample of 80 children uniformly across the three selected centers — with a goal of about 27 children in each center (about 16 of whom would be assigned to the treatment group, and about 11 of whom would be assigned to the control group at each center). For small centers this may impose too great a burden in terms of the need to recruit an additional 11 children for the control group — or even to identify 16 participants to include in the treatment group for the smallest cases. For example, according to the latest PIR data, the average (and median) Head Start center serves about 50 children (an average of 458 children in an average of 9 centers). As a consequence, our fixed sample of about 11 additional children for the control group would represent an increase of about 22 percent for such a program. For some center directors this may represent an unreasonable burden.

We plan to deal with this issue once we have selected our sample of grantees/delegate agencies by working with the local program staff to devise a plan for how to allocate our desired sample of study participants among available centers while, at the same time, considering the implications for data collection costs. In some grantees/delegate agencies we may be able to obtain our desired sample from three (or fewer) centers; while in other situations we may have to expand the sample of centers necessary to obtain the necessary child sample. The only requirement is that the selection of centers must be done at random with known probabilities from those "eligible" to be included in the study. We cannot simply draw a sample of centers and then sort out and adjust for any problems with numbers after the fact.

---

[26] As noted above, extremely new centers will not be considered eligible for inclusion in the study, and we will also have to check on the extent of "saturation" at the center level to ensure that we will be able to create a control group.

[27] The average number of centers per grantee/delegate agency is 9.4 and the median is 6.0. The range is from 1 to 166 centers.

To help illustrate this plan, consider an "average" grantee/delegate agency serving about 460 children in eight Head Start centers — four of the centers serve 50 children each (a total of 200), two serve 100 children each (a total of 200), and there are two small centers each serving 30 children (a total of 60 children). If we were to systematically select three centers (with probabilities proportional to size) we could end up with one center of each of the three sizes, i.e., one at 30, one at 50, and one at 100 children. The added recruitment of 11 children for the control group would, therefore, represent 37%, 22%, and 11% of current enrollment at the three centers, respectively. If this turned out to be a problem for the smaller center, we could, for example, re-allocate the control group children to the three centers using a rule of proportionality i.e., sample 5 children from the 30-enrollment center, 9 from the 50-child center, and the remaining 18 from the 100-child center, figures equal to 18% of current enrollment in every case.[28] Or, alternatively, we could select a sample of four centers from this grantee and allocate the sample either on a fixed or proportional basis across the four rather than three centers.

Similar problems may be encountered in situations where there are *disproportionate distributions of 3- and 4-year olds across centers* operated by a selected grantee/delegate agency. In these circumstances, we will follow the same general strategy of working with grantee/delegate agency staff to devise an acceptable strategy for allocating our desired total sample of 3-year-olds and our desired total sample of 4-year-olds among centers.

Why don't we vary the sample size across grantees/delegate agencies?

Given the variation in enrollment across centers—and, at the prior level of sampling—across grantees/delegate agencies, one seemingly simple solution would be to allocate samples of Head Start children to the selected grantees/delegate agencies proportional to *their* overall enrollment size, i.e., to select larger child samples from larger grantees/delegate agencies and smaller samples from smaller programs. We have instead chosen to select a ***fixed*** sample of 80 newly-entering children from each sampled grantee/delegate agency for several reasons.

First, allocating equal samples to each grantee/delegate agency provides greater control over the size of the total national sample and the size of the samples in the various analytic strata of importance. If the sampling rule for children was proportional to the size of the selected (and recruited!) grantees/delegate agencies, the size of the final sample of Head Start children for the nation or any strata would be driven by the sizes of the recruited sample of programs (along with the maximum sampling rate each program could sustain, particularly as regards the control group). This could yield too few children for the intended analyses if the set of grantees and delegate agencies selected happened to be below average in size.[29] The use of a fixed sample size of children, as we have planned, provides the best method of ensuring that we have a sample of children in each strata and the nation that can support the intended objectives of the impact analysis.

---

[28] We arrive at these figures by first noting that the goal is 32 control group children for each sampled grantee/delegate agency (rounding to the 11 per center when divided evenly). These 32 children are to be distributed proportionately among three centers that collectively serve 180 children (30+50+100), a ratio of .178 control children for every child now served (32/180=.178). This translates into a control groups at each center equal to 17.8 percent of its existing enrollment.

Also, with samples set proportionate to grantee size, the sample of children becomes more heavily concentrated in a small number of relatively large grantees/delegate agencies. This increases the design effect caused by clustering and reduces the precision of all estimates. The same shift affects the number of children observed under different Head Start program characteristics and community contexts, factors we want to vary and examine for analysis purposes. Thus, concentrating a greater share of our observations in a few settings (those of larger grantees/delegate agencies) would reduce our ability to address the "what works for whom?" question that constitutes the second major goal of the study.

Step 9: Divide the Population of Head Start Children into One-Year and Two-year Participants

The original legislative mandate required that the Head Start Impact Study "…to the extent practicable, consider addressing possible sources of variation in impact of Head Start grantees/delegate agencies, including ….. the length of time a child attends a Head Start program; the age of the child on entering the Head Start program …. " To respond to this requirement, we plan to include roughly equal samples of **newly entering** 3-year-olds (who will be studied through two years of Head Start participation) and **newly-entering** 4-year-olds (who will be studied through one year of Head Start participation). We have focused the study on **newly-entering** children so as to isolate the effects of Head Start apart from any prior exposure to the program's benefits[30]. We also planned on an equal sampling of 3- and 4-year old enrollees despite the fact that 4-year-olds represent about twice the proportion of all Head Start participants than do 3-year-olds. In large part, this is because the 4-year-olds include both newly entering 4-year-olds plus returning children who began Head Start as 3-year-olds and who have turned 4 years of age in their second year of program participation.

It is recommend that there be separate representation of these two age cohorts for two primary reasons. First, it is expected that there will be very different program impacts associated with one year versus two years of Head Start experience. In other words, duration of treatment should matter a great deal to program impacts, especially when one considers that by the time a child is 5 years of age, two years of Head Start represents 40 percent of his or her total life span compared with 20 percent for those entering Head Start at 4 years of age.

---

[29] Above-average grantees/delegate agency selection could be offset by a uniform reduction in the child sampling rate across all selected grantees/delegate agencies.

[30] Children previously enrolled in an Early Head Start program will be excluded from random assignment because they are, in most cases, ensured of continuing participation in regular Head Start.

Second, there are important shifts occurring in the age composition of Head Start nationally that are likely to have important future policy consequences. With the growth in the availability of alternative preschool options, there is some anecdotal information suggesting that grantees/delegate agencies are serving increasing numbers of the eligible 4-year-olds. At the same time, Head Start grantees are increasing their enrollment of 3-year-olds, and there is the obvious increase in younger children as Early Head Start grantees/delegate agencies become more prevalent. As a consequence, there is growing relevance for exploring the impact of Head Start given two years of participation rather than one.

The bottom line is that if the interest of the study were primarily in children as a whole, then the most efficient sampling procedure would be equal **sampling rates** for each type of child. But because we believe that the main interest is in looking separately at children with one year and with two years of Head Start services, the optimal sampling procedure is for equal **sample sizes** for each group.

Step 10: Select Appropriately-Sized Samples of Head Start Children

In the selected 225 Head Start centers, spread across up to 75 study grantees/delegate agencies, we propose to select an **initial sample** of 3,137 newly entering 3-year-old participants and 2,541 newly entering 4-year-old participants. As shown in Exhibit 2, a total of 1,882 3-year-olds will be assigned to the treatment group and 1,255 to the control group, while a total of 1,524 4-year-olds will be assigned to the treatment group and 1,017 4-year-olds to the control group.[31]

---

[31] To the extent that the final study sample includes less than 70 grantees/delegate agencies, the average samples of Head Start children selected in the remaining sites will be increased to preserve the size of the desired overall sample.

Exhibit 2: Expected Sample Size At Each Wave Of Data Collection

COHORT 1: TWO-YEAR PARTICIPANTS (3-YEAR-OLDS)

	70 Participating Grantees/Delegate Agencies		
	Treatment	Control	Total
At Random Assignment	1,882	1,255	3,137
Fall 2002 HS	1,694	1,130	2,824
Spring 2003 HS	1,524	1,017	2,541
Fall 2003 HS	1,372	915	2,287
Spring 2004 HS	1,235	823	2,058
Spring 2005 K	1,111	741	1,852
Spring 2006 1st grade	1,000	667	1,667

COHORT 2: ONE-YEAR PARTICIPANTS (4-YEAR-OLDS)[1]

	70 Participating Grantees/Delegate Agencies		
	Treatment	Control	Total
At Random Assignment	1,524	1,017	2,541
Fall 2002 HS	1,372	915	2,287
Spring 2003 HS	1,235	823	2,058
Spring 2004 K	1,111	741	1,852
Spring 2005 1st grade	1,000	667	1,667

[1] Includes an assumed 10% attrition rate each year.

Exhibit 2 also indicates the anticipated sample sizes for each wave of data collection (see the Measurement and Data Collection Plan). To obtain a final sample of 1,667 three-year-olds and 1,667 four-year-olds at the end of the study period—the size needed for adequate statistical precision (see the next section)—we estimate a beginning sample size of 3,137 3-year-olds and 2,541 newly enrolled 4-year-olds. On average, this will mean sampling approximately 45 3-year-olds and 36 4-year-olds from each of the anticipated 70 participating grantees/delegate agencies.

Within the sampled Head Start centers we are ***not*** planning to sub-sample classrooms and then confine the children taken into the sample to those classrooms. Instead, the sample

BEST COPY AVAILABLE

of children will be distributed across all classrooms at random, according to the natural patterns that emerge as children are assigned to classrooms by the center director and her/his staff. However, *estimates of the influence of classroom factors on child outcomes and impacts* will, as described in a later section, still be possible under this design using the proposed HLM analysis framework, as long as classroom characteristics are measured and given their own level in the model's hierarchy.

2.2 Sampling Issues

Sample Attrition

Built into each wave of data collection is an assumed 10-percent attrition rate. This attrition rate is based on Westat's experience with the Head Start Family and Child Experiences Survey (FACES) and includes refusals, children/families who moved from the area (and drop out of Head Start), and children and families who could not be located. By the third wave of FACES data collection (the timing of waves is comparable across studies), parent, teacher, or child assessment data were collected on 69 percent of the original sample, for an average of 10-percent attrition in each wave. Based on our FACES experience, we expect a greater part of the attrition to be a result of children dropping out of the Head Start program and moving from the area than a result of outright refusal. These assumptions are also supported by the results of other studies of similar populations (e.g., the Comprehensive Child Development Program evaluation).

Although we will strive to achieve lower rates of attrition, we have taken a more conservative approach to sampling and planning to minimize the risk of having insufficient sample cases to support the analyses throughout all waves of data collection. That is, to assume better circumstances could seriously undermine the overall study if one's high expectations prove to be unrealistic. If our more conservative assumptions are wrong, the precision of the estimates will be increased.[32]

Unequal Allocation to Treatment and Control Groups

As noted in Exhibit 2, the random assignment sample is not divided equally between the treatment (Head Start participant) and control groups. Instead, we plan to allocate 60 percent (3,406 of 5,678 initial interview attempts) to the treatment group, and 40 percent (2,272 of 5,978 initial interview attempts) to the control group. This imbalance in the randomized sample reduces the precision of the impact estimates by just 2 percent, compared to a balanced 50-50 design, and saves considerably on data collection costs (because treatment group members—who participate in Head Start—require less effort to track and interview over time than control group members). It also reduces the number of control group members (i.e., additional unserved children to be recruited) required of each site for a given total sample size. This second point expands our ability to identify "eligible" grantees/delegate agencies, and minimizes the burden on the study sites.

---

[32] Of course, the added sample does have budgetary implications that are the cost of reducing the undesirable down side risk. Fortunately, this can be handled once the full-scale study is underway—if ACYF desires—by randomly reducing study samples if they exceed expectations.

Uneven random assignment ratios such as the 60-40 ratio proposed here are not uncommon in experimental research[33] and in no way undermine the assurance of internally valid (i.e., unbiased) impact estimates that makes randomized designs so attractive in the first place.

Excluding Very High-Risk Children

The selected grantees/delegate agencies will be allowed to exclude a **limited** number of very high-risk children. This decision was made for three reasons: (1) we have serious ethical concerns about assigning very high-risk children to the control group, especially in situations where Head Start may provide their only option for early childhood services; (2) our past experience with the use of random assignment for vulnerable populations has demonstrated that this issue can often be a "deal breaker" when trying to recruit study sites; and (3) there are some children who cannot be assigned to the control group because they have been placed in Head Start by the local child welfare agency (or, more correctly, by the jurisdictional court) as part of their foster care placement and must, by law, receive Head Start services. The definition of "very high-risk" will be made on a case-by-case basis with each grantee/delegate agency and in close consultation with ACYF staff. Examples of exclusions that may be allowed include: children of homeless families, children in families with documented abuse and neglect, and children with severe disabilities, especially those that would prevent them from being tested.

Placing 3-Year Olds Into The Control Group For Two Years

Our plan is to randomly assign both 3- and 4-year olds at the time they initially apply for entry into Head Start. In both cases, those children who are assigned to the control group would be excluded from Head Start — for the 4-year olds, this is a one-year exclusion, while for the 3-year olds the exclusion is for two years until they reach kindergarten age. Some grantees/delegate agencies may be reluctant to exclude newly entering 3-year olds who are randomly assigned to the control group for this period of time, i.e., the loss of two years of comprehensive services may be viewed by some grantees/delegate agencies as too high a price to pay for the sake of rigorous impact evaluation. Yet, to measure the full effects of the program on two-year participants, we need to contrast children who receive two years of services with control group members who receive non-Head Start services. Moreover, answering the question of two-year effects is critical from the standpoint of both early childhood development and the policy and fiscal issues surrounding Head Start.[34]

---

[33] The National JTPA Study used a 67-33 random assignment ratio, for example.

[34] A reasonable alternative to multi-year exclusions from Head Start would be to focus the study not on the overall value of two years of participation, but on the value of the *marginal* year of participation—what the early (i.e., age 3) year ***adds*** to the impact of the standard, single (i.e., age 4) year. This could be accomplished by randomly assigning newly-entering 3-year olds to a one-year control group and comparing their outcomes to those of full two-year participants in the treatment group. This design could be extended to approximate the effect of two years of participation compared to none without actually excluding any children for two years, using a methodology that provides upper and lower bounds on the true effect. This extension would also allow us to assess how much a single year of enrollment would benefit the children currently served on a two-year track. A variation on our basic experimental design that incorporates this idea is provided in Appendix D.

Families (or Households) With Multiple Eligible Children

Several possibilities involving siblings, and unrelated household members, can also complicate random assignment, including:

- ***Twins*** — in most cases, parents will probably apply to Head Start for both children at the same time. Because we want to avoid having children in the same family assigned to different study conditions, we would randomly assign **both** children to either the treatment or control group. If both children do not seek enrollment at the same time, then the non-applying child would fall into one of the four following conditions.

- ***Multiple Newly-applying Children Of Different Ages*** — for example, a parent may apply for enrollment of a 3- and a 4-year old child at the same time. Again, because we want to avoid multiple children in the same family being assigned to different conditions, we would randomly assign both children to the same group.

- ***Other Non-sibling Children In The Same Household*** — a newly applying child may have unrelated, but eligible, children living in the same household who may apply at the same or at different time. As above, we would want to randomly assign the multiple children to the same study condition.

- ***Younger Sibling(s) Who Can Subsequently Apply For Head Start*** — for example, a newly-applying 3- or 4-year old could have a younger sibling at home who could apply to Head Start in a subsequent year. Ideally, we would want to randomly assign both (or multiple) children to the same study condition. However, this may be logistically infeasible and is an issue that we will have to explore with the selected grantees/delegate agencies. If we elect to ignore subsequent enrollees, we should have only relatively minor contamination of our study groups, and this is something that can be determined as part of the planned data collection activities and taken into account during the subsequent analysis.

- ***Currently or Previously Enrolled Sibling*** — the last example is a situation in which the "target" child has either a currently enrolled sibling or one or more siblings who have been previously enrolled in Head Start. In this situation, if the target child is assigned to the treatment group the program's impact for that child may reflect services received by the parents in connection with the other child or "spillover" effects from services delivered to the sibling. Alternatively, if the target child were assigned to the control group, he or she would be affected by Head Start services.

Most of these types of situations will be handled by treating the family/household — rather than the "target" child — as the unit of random assignment, i.e., all children currently applying for Head Start would be assigned to the same study group (either treatment or control). The last example is clearly more complicated. We could, for example, screen out of the study any child whose current custodial parent(s) or guardian(s) have previously had — or currently have — a child in Head Start since we cannot represent such children in the control group in an "unaffected" state. But this may not be the best solution because parents (or Head Start staff) could discover that reporting a prior child in Head Start is a convenient way to assure the target child is not put into the control group. More importantly, we would be significantly changing the population of children served by Head Start who are included in the study. The same problems arise if we simply put all such children into the treatment group, with the added disadvantage of creating an imbalance in the types of families compared across the treatment and control groups. Instead, we plan to include such children in the study — accepting whatever spillover effect my be present — and to collect information on older sibling participation in the parent baseline interview. We will then use these data in our analysis to gain some leverage on the "transfer" effect of the parent component of Head Start between siblings.

Grantees/Delegate Agencies Do Not All Serve 3-Year Olds

There will be a few instances where the selected grantee/delegate agency will be unable to meet our requirements for equal samples of 3- and 4-year old children. Based on recent PIR data, about 2.4 percent of all grantees/delegate agencies serve no 3-year old children, and about another four percent serve 30 or fewer 3-year olds **and** 30 or fewer 4-year olds. Our plan to combine "small" grantees/delegate agencies (see above) will certainly help, but it is likely that we will, albeit rarely, have grantees/delegate agencies in the study sample that will be unable to meet our sample requirements for newly entering 3-year olds. In such cases, we will adjust our sample sizes in other sites to offset the reduction, as discussed in connection with Step 6 of our sampling plan above.

2.3 Statistical Power

The standard measure of the confidence that one can place in estimated program impacts is called the ***minimum detectable effect (MDE)***, i.e., the smallest true impact one would expect to identify as statistically significant. This refers to program effects that can be detected as "statistically significant" 4 out of every 5 times they occur, a level of confidence known as "80-percent power." There are several types of outcome differences that are of interest in this study:

- National impact estimates for the full population, based on comparisons of the treatment and control groups (e.g., PPVT scores for treatment group minus PPVT scores for control group).
- Impact estimates for subgroups of the population, based on comparisons of subsamples of the treatment and control groups (e.g., a subgroup could be comprised of children from areas in which there is a high penetration of Head-Start-like alternative programs for low-income children.)
- Comparison of impacts between contrasting subgroups (e.g., an impact estimate for girls minus an impact estimate for boys.

Exhibit 3A provides estimated MDEs for the first two types of comparisons, while Exhibit 3B addresses the last category. The figures given are for a simple random samples and do not take into account the effect of sample clustering (e.g., drawing samples of children from the same geographic cluster or grantee/delegate agency).[35] Exhibit 4 (provided below) illustrates the effect of this clustering on MDEs.

In these exhibits, we consider several different outcome measures, but the Peabody Vocabulary Test (PPVT) is particularly good as the "test case" outcome for two reasons: among the outcomes of interest to this study, it has one of the highest variances—making it a conservative test of the power of the sample for outcomes generally—and it provides one of the broadest and most central indications of school readiness and social competence available to the study. The national average score for this test is 94 (including both Head Start and non-Head Start children).[36]

Exhibit 3A shows minimum detectable effects, or MDEs, for the national impact analysis; these figures represent the smallest true effect of Head Start on the average national participant that will be detected with 80-percent power using the proposed sample sizes. Here, we look at MDEs for the PPVT, a social awareness index, and two different population percentages: 50% and 30%. The first column gives MDEs for comparisons of treatment and control based on the entire sample size of 1,677 (1,000 treatment and 667 control). Thus, for example, the MDE for the PPVT based on the entire sample is 1.83. The second column gives MDEs based on half the sample size, or 833. This would be applicable, for example, if one were considering the PPVT scores of boys. The third column gives MDEs based on one-fifth of the sample size, or 333. Finally, the fourth column gives MDDs based on one-tenth of the sample size, or 167.

Exhibit 3A: Minimum Detectable Effects for 3- or 4-Year-Olds

Head Start Participants (end of 1st grade)

Child or family outcome measure	Population mean[1]	Minimum detectable difference for subgroup sample size			
		N=1667	N=833	N=333	N=167
Peabody Vocabulary Test (PPVT)	94	1.83	2.58	4.10	5.78
Social Awareness Index	4.06	0.20	0.28	0.45	0.63
Treatment percentage = 50%	N/A	7.0%	9.9%	15.5%	21.5%
=30%	N/A	6.6%	9.4%	15.2%	21.5%

[1] Population means and standard deviations (not shown; = 13 for PPVT and = 1.42 for the social awareness scale) come from FACES 1999 data.

---

[35] The figures in Exhibits 3A and 3B apply to impacts on children and families at the end of the first grade year. Even **smaller differences** in impact will be detectable with 80-percent power at the end of the kindergarten year and at earlier points (due to lower data collection attrition and, hence, larger sample sizes). Because they are based on equal-sized samples, the estimates would be the same for both 3-and 4-year-old Head Start participants.

[36] This figure, and other data used in Exhibits 3 and 4, come from FACES 1999 data.

BEST COPY AVAILABLE

For the full population, the design is able to detect quite small effects, effects of fewer than two points on the PPVT scale. Analysis of the FACES data shows an average gain in PPVT score of 4 points over national norms between fall and spring testing of Head Start children. A program impact based on a true control group could differ from this trend-based approximation. Even so, an impact of 3 or 4 points is not out of the question, while one of 2 or 3 points would be worth knowing about and possibly signal a successful program. This suggests that an MDE of 2 PPVT points is an appropriate level of investment in sample size for the national evaluation. It seems particularly adequate given that we expect to be able to detect even smaller proportionate effects on other, less variable child and family outcomes. Three other outcomes are illustrated in the exhibit: the Social Awareness Index (SAI) and 50% and 30% characteristics of the population such as up-to-date immunization. For example, we will be able to detect quite small changes in the SAI as a percent of the mean.

Exhibit 3B provides estimated MDDIEs for the "differences of differences," e.g., a PPVT score for boys in the treatment group minus that for boys in the control group, **minus** that for treatment group girls minus that for control group girls. The first column gives estimates for equal sized subgroups, such as boys and girls. The other columns give estimates for unequal sized subgroups: 60% compared to 40%, 75% compared to 25%, and 90% compared to 10%.

Exhibit 3B: Estimated Minimum Detectable Differences In Effect, Comparison of Two Subgroups of 3- or 4-Year Olds

Head Start Participants (end of 1st grade)—Peabody Vocabulary Test

Child or family outcome measure	Population mean[1]	Minimum detectable difference in effect[2] by ratio of subgroups 50 / 50	60 / 40	75 / 25	90 / 10
Peabody Vocabulary Test (PPVT)	94	3.6	3.7	4.2	6.1
Social Awareness Scale	4.06	0.41	0.42	0.49	0.70

[1] The population mean and standard deviation (not shown; = 13) for PPVT and = 1.42 for the social awareness scale) come from FACES 1999 data.

[2] Based on a two-tailed t-test of the statistical significance of a difference-in-difference estimator (average outcome for subgroup A, treatment minus control, *minus* average outcome for subgroup B, treatment minus control). The confidence level of the test is set at 95 percent (significance level = .05) and the degree of confidence required in detecting a true difference in impact (i.e., the power of the test) is 80 percent.

As can be seen, the proposed sample design can detect a 3.6-point difference in impact on the PPVT between two equal-sized subgroups of 3- or 4-year-old participants. For example, if Head Start increases the average score for male participants from 88 to 93 and for female participants from 90 to 91, we would expect to identify these two

BEST COPY AVAILABLE

effects—5 scale points and 1 scale point—as different from one another and hence conclude that Head Start works better for boys than for girls. A true difference in effectiveness of 3.6 points or less might not be detected as statistically significant in the analysis, although it often still would be, in fact, the case.[37] Similar results hold for subgroups of unequal size, up to a 75/25 split of the population. In comparing subgroups of very unequal size, such as the 90/10 ratio shown in the last column, impacts would have to differ by 6 scale points for us to have an 80 percent chance of finding a statistically significant difference.[38]

Exhibit 3B also shows that the sample could detect a 0.41 difference in impact between two equal-sized subgroups on a Social Awareness Scale that has been used in FACES. This is a somewhat larger relative difference (0.41/4.06) than for PPVT (3.6/94), due to there being more relative variation among 3 and 4 year olds in this scale than on the PPVT. If, for example, Head Start increases the average scale score for male participants from 4.0 to 5.0 and for female participants from 4.0 to 4.5, we would expect to identify these two effects as different from one another and hence conclude that Head Start works better for boys than for girls. We could also expect to identify such effects as different for a 75/25 subgroup split, but would not necessarily be able to detect such effects as different for a 90/10 subgroup split.

The calculations are, in fact, a bit more complicated than has been indicated thus far, since they depend not just on overall sample size but on the extent to which sampled cases "cluster" in certain Head Start centers or under certain Head Start grantees/delegate agencies. Clustering more observations in a geographic region or organization than would occur naturally in a simple random sample creates what are called "design effects" for the study. These effects reduce the precision of all estimates derived from the sample. The lost precision—and consequent increase in standard errors and MDEs—depends on two factors: how much the average "cell size," or number of observations per unit, exceeds the size expected through purely random sampling; and the degree to which cases in the same cell experience similar impacts, a factor known to statisticians as the "intraclass correlation" in impacts. Design effects from cells only slightly in excess of their expected sample size under purely random selection can safely be ignored.

This is the case when selecting grantees/delegate agencies from just 25 of the approximately 170 GGCs in the nation. We will select three grantees per GGC, which is only modestly above the expected number of 1. It also applies in selecting an average of 3.2 Head Start centers per grantee/delegate agency, a level scarcely above the 2.5 centers expected with simple random selection.

---

[37] A 3-point difference in impact, for example, might be detected as statistically significant only 60 percent of the time.

[38] Actual MDDIEs would be slightly smaller than those shown when subgroups are analyzed using an HLM model, for example, that predict outcomes as a function of treatment, subgroup (independently and interacted with treatment), and various baseline characteristics. Precision is gained by including background variables in the model--participant, environmental, and programmatic baseline measures that also influence outcomes. If the inclusion of these variables increases the explanatory power of the model from 5 percent (regression R-squared of .05) to 15 percent (regression R-squared of .15), for example, MDDIEs would decline by 5 percent. Here, the MDDIE for equal-sized subgroups would fall from 3.6 to 3.4, and similarly for other findings in the exhibit. A very large--and unrealistic--increase in the explanatory power of the model would have to occur for the general tenor of the results to change appreciably.

Design effects are potentially important at the final level of the sampling plan, when selecting Head Start participants at random from the sample of 225 Head Start centers selected. If participants experiencing relatively large impacts from Head Start services tend to cluster in one subset of Head Start centers, while those experiencing smaller than average effects cluster in another, picking just some of the centers (225 of 9,523) will increase the variance of the sample estimate as well as the minimum detectable effects for the analysis.[39]

The columns of Exhibit 4 trace out the consequences of design effects due to clustering in the proposed design.[40] With a small intraclass correlation of .10 or less, clustering by center does little to MDEs, which remain near 2 scale points for the Peabody Picture Vocabulary Test (PPVT) score. Larger correlations can as much as double MDEs (for any outcome, including the PPVT score), making it less likely that relatively small effects will show up in the analysis as statistically significant. This raises two questions: (1) How much correlation in impacts should we expect within Head Start centers? and (2) How small an effect do we need to be able to detect to appropriately assess Head Start's contribution nationally? Neither has a sharp answer, though both can be approached productively and—in this case—suggest that the proposed sample very likely will detect any Head Start impacts of national importance that do occur.

Exhibit 4: Estimated Minimum Detectable Effects With Clustering

Head Start Participants (end of 1st grade)—National Estimates

Child or family outcome measure	Population mean[1]	Minimum detectable effect[2] by clustering within centers (intraclass correlation)					
		0	.05	.10	.20	.30	.50
Peabody Vocabulary Test (PPVT)	94	1.8	2.0	2.1	2.4	2.6	3.0

---

[39] The corollary to this point for the differential impact analysis discussed earlier has second order importance and thus does not require assessment of MDDIEs. For design effects to arise there, Head Start participants must cluster together in centers based on how sensitive their impacts are to different baseline factors. Thus, children from backgrounds where, say, gender strongly influences how much can be gained from Head Start participation would have to cluster together in one set of centers while those whose gains are relatively insensitive to gender would have to concentrate in other centers. This pattern--and its equivalent concerning children whose impacts are more or less sensitive to variations in program features and/or environmental factors--seems unlikely.

[40] The clustering adjustment is based on an average cell size of 4.4, the number of completed treatment interviews expected per center in the final round of data collection with 10 percent attrition per round, for each of the separate 1-year and 2-year participant samples. As noted earlier, calculating impacts using a model that includes descriptive variables about participants, grantee/delegate agencies, and environments will reduce standard errors and MDEs relative to those shown here, to some extent offsetting the upward influence of the design effects illustrated. Unfortunately (as also discussed previously) this offset is likely to be quite small.

BEST COPY AVAILABLE

PPVT score *levels* correlate strongly within centers nationally, with an intraclass correlation of .51.[41] However, this does not mean that the degree to which Head Start improves test scores—an *impact* measure, not a level—will associate so strongly between children in the same Head Start centers. An intraclass correlation of .20 or .30 for impacts seems more reasonable, and it might be much lower.[42] (Since this is the first impact evaluation of Head Start to use random assignment, no reliable information exists for calculating intraclass impact correlation empirically.) Thus, it seems reasonable to assume that the proposed sample will be able to detect PPVT effects at least as small as 3 points taking account of clustering, and probably as small as 2 points.

Why Do We Need Such A Large Sample?

As noted above, Head Start need only increase PPVT scores nationally by 2 to 3 scale points for its benefits to be **firmly** established under our proposed research design. To see the importance of this assurance, imagine a study of lesser size and the outcome the Head Start community might face in that instance. The world—and Congress—is likely to see this evaluation as a full test of the pay-off to school-readiness assistance to disadvantaged preschoolers, or at least as the final word on Head Start as a service delivery mechanism. **Our fear is that this will be the case regardless of how weak the study becomes in terms of sample size.** An evaluation incapable of finding (or at least statistically very unlikely to find) two to three point gains in PPVT scores because of Head Start will be taken as evidence that Head Start does not work. A contrasting study, with samples set adequately at the level proposed here, would quite likely enable an adequate test of the question regarding whether Head Start moves children forward to prepare them to learn.

Ironically, from a policy standpoint, a study of insufficient size could prove worse than no study at all. As the Department's "best shot" at objectively assessing Head Start's achievements on a widely accepted basis—and of finding ways to improve the program for the future—the risk of reaching the wrong conclusion because of insufficient investment in sample points seems to us to more than justify the level of effort proposed in our design. While examination of other key indicators of program success might suggest somewhat different balance points on sample size than the PPVT test score used here, the principals will not change and the conclusion that sample sizes be maintained at current levels will likely ring through as clearly as ever. While modest cuts below proposed levels would not drastically reduce the reliability of the study, we see no reason to run the risk having invested in so many other ways in the best possible tool for examining and improving Head Start's services to low-income families and children.

---

[41] This figure is derived from the FACES data. This is an exceptionally large intraclass correlation, but it is not surprising that higher achieving children tend to group together in centers, given the importance of socio-economic background and other family factors in influencing both the pace of child development and residential location.

[42] There is no particular reason to expect children to cluster within center based on their ability to gain from Head Start services, as opposed to how well they may score overall. The socioeconomic and family factors that associate with place of residence and hence lead to strong clustering of test level by center act on the portion of test scores determined by home environment, not the portion determined by pre-school program (the nexus of any Head Start effects).

Will The Sample Remain Representative Over Time?

Essentially all surveys have some degree of nonresponse. In general, nonrespondents will not be identical to respondents in some respects, which leads to the likelihood of bias in survey estimates. For this reason, it is important to keep nonresponse to a minimum, and we have planned an extensive effort to contain nonresponse.

Despite our best efforts, however, after several rounds of data collection, the level of cumulative loss of participating children may be fairly high. It would be irresponsible to claim that children who are not included in the survey will be very similar to respondents across the full range of data to be collected. No one can know in advance in what ways, and by how much, non-respondents may differ from respondents.

As a consequence, as described in the Analysis Plan we intend to employ a methodology that will reduce the biasing effects of nonresponse. We anticipate obtaining data on about 90 percent of all sampled children in the initial interview period. This will provide the full set of data collected initially for all of these children, including direct assessment scores. This will then allow us to determine categories of respondents and nonrespondents with similar test scores and other characteristics and make a weighting-class nonresponse adjustment separately for each category. Because such data will correlate highly with future responses, the nonresponse adjustment procedure should be quite effective in reducing the bias due to nonresponse. This is a much better situation than occurs for most surveys, where typically very little information is available for nonrespondents.

As an alternative to the weighting class nonresponse adjustment described above, we will also consider using a *response propensity* nonresponse adjustment procedure. In this procedure, a logistic regression algorithm is used to determine the predicted probability of response. This procedure has the potential to be more effective in reducing nonresponse bias than a weighting-class procedure when there is data available on a large number of characteristics for nonrespondents, as is the case for this study.

We also propose to produce tables that compare the characteristics of respondents, nonrespondents after the first data collection, and all children (including nonrespondents) on a range of characteristics obtained in the first data collection period. To the extent that nonrespondents and respondents are not too dissimilar on most characteristics, this will reassure readers of the final reports that the sample of respondents is representative of all Head Start children.

3. Field Test Plan

3.1 Introduction

Prior to the start of full-scale random assignment and data collection in the Summer and Fall of 2002, a field test of all study procedures will be conducted, starting in early **Spring of 2001**. This test will include collection of all program, parent, and child data, and in particular a "trial run" of the proposed procedures for recruiting sites, as well as, conducting and monitoring random assignment in the **Summer of 2001**. Information derived from the field test will be critical for further shaping the design and implementation of the main study prior to full-scale implementation in the **Summer of 2002.**

Data collection for the field test must be started by the **Fall 2001** to inform the plans for the full-scale study. As a consequence, all field test sites will be recruited, and application procedures and parental notification, put in place by **Spring 2001**. It is important to keep in mind that any field test is going to be somewhat limited in terms of what it can teach us, i.e., for many of the potential things that we could learn about, a "small" field test is unlikely to be large enough to really protect us against all of the things that can go wrong in the full-scale study. Therefore, we expect that our recruitment of the main study sample will also serve to inform our procedures of gaining cooperation and participation; tailoring random assignment procedures; identifying agency eligibility and enrollment differences; and finding solutions to issues that could potentially affect successful implementation of random assignment with each of the sampled 75 grantees. Multiple site visits by recruitment teams and follow-up visits by site coordinators will be conducted. Individual plans will be created for each sampled site and a Memorandum of Understanding developed, outlining the plans for identifying children, random assignment, and data collection.

Purpose Of the Field test

The primary purpose of the field test will be to test the workability of **site recruitment** and **random assignment** procedures across a small number of communities, and collect comparable data from both the Head Start and non-Head Start control group. Because the current plans call for only marginal changes to data collection instruments that have already been well tested in other studies (e.g., FACES, ECLS-K), almost any set of sites for the field test will provide a needed test of the existing set of measures, questionnaires and other protocols.

What, then, are the most critical things that can be learned from such a field test? Although there are many answers to this question, below are the issues that are believed to be most important:

- Learn the degree to which **saturation** may exist across different communities and the degree to which more recent changes in HS-partnerships with other providers and other "blending" options have increased the number of saturated communities.

- Learn more about the factors that may affect grantee/delegate agency decisions to participate in the study, and how we can work to maximize **program cooperation** and **participation,** without disruptions to program operations or having any undue burden being placed on staff. This information could help to increase the "external validity" of the full-scale study. In particular we will test how different conditions affect the ease or difficulty of grantees to provide 32 control cases.

- Learn how the **proposed random assignment procedures** will work in a range of Head Start programs and communities to support grantee's participation in the study and minimize or eliminate any difficulties with implementing random assignment that may potentially compromise sample integrity and/or diminish data quality.

- Learn whether or not normal program **recruitment strategies** and **program operations** are substantially altered by the necessity to recruit a small number of additional families for the control group. In other words, does the additional recruitment effort result in a substantially different composition of children and families normally served?

- Learn about what happens to children and families who are assigned to the **control group** and end up utilizing a wide range of available care options in their communities. Little, if anything, is known about what families assigned to the control group might do in terms of finding alternative arrangements for their child. This information would better prepare us for likely consequences (as well as points of difficulty in implementing random assignment at the program level), and have the added benefit of providing policy-relevant information about the range and quality of preschool and childcare experiences of low-income children who lack the opportunity to enter Head Start. (The latter information would be a useful addition to the 2003 report to Congress.)

- Learn better ways to avoid control group contamination, particularly control group members "crossing over" to the treatment group by subsequently enrolling the following year, or in another Head Start program. This information could increase the reliability of the impact estimate in some sites.

- Develop better data collection designs and/or procedures for the full-scale study, without changing the *content* of the data collection instruments. These adjustments could increase overall data quality. In particular, testing the ability of one site coordinator to adequately serve three geographically clustered grantees.

- Based on the combination of some or all of the above factors, learn more about the **practical implementation issues of a national randomized study design,** as well as specific situations in which it may not be possible to implement randomization successfully. This could avoid an investment in a failed full-scale implementation (i.e., either no useable study at all or one with questionable representativeness), and provide an opportunity to look for alternative design strategies.

3.2. Field Test Design

The field test design will take a limited number of sites from recruitment through data collection to test those factors that are both measurable and likely to affect the success of the randomized study. (The first criterion is necessary because there are many things that can pose difficulties (e.g., lengthy negotiations with program directors or various staff) that are nearly impossible to predict in advance and which can be used to select field test sites. There are probably a range of factors that would meet this criteria. The following are ones that we believe are most important to test because they are likely to have significant implications for our success, and testable with a reasonably- sized field study: We need the experience of trying out study procedures across a range of circumstances that pose different challenges, so that we have the right breadth of experience when working with main study sites. Five factors that could lead to different challenges are measurable in advance to guide field site selection:

- **Local Context** — the extent to which the program operates with a "service rich vs. a service poor" environment is expected to affect both the willingness of staff to agree to, and maintain the integrity of, random assignment. It will also affect the experiences of the children who are assigned to the control group.

- **Saturation** — the extent to which there are unserved children in the community is also likely to have a strong influence on the willingness of staff to agree to, and maintain the integrity of, random assignment.

- **Auspice**---whether the grantee/delegate agency is within a school system or some other administrative entity may require different recruitment strategies.

- **Urban location**—the urban, suburban, or rural status of a grantee/delegate agency will present different challenges to implementing the study. Varying these contexts will allow us to test how the different challenges these programs may face such as recruiting families or transportation issues will influence the programs' willingness and ability to participate

- **Grantee/delegate agency size**—the smaller grantees may have more difficulty identifying sufficient numbers of children for a control group and perceive the study as more of a burden to local program staff. Large grantees will be defined as those with 450+ children and small grantees will be those with under 450 children[43]

In addition, the design of the field test needs to take into consideration the current budget as well as the time available to recruit and negotiate with the field test sites, and to implement enrollment and random assignment procedures by the **Spring of 2001.**

---

[43] The average grantee size is 450 children.

Given these various factors, the following is our suggestion of how these design parameters could play out in a field test:

- Select three (3) geographic clusters that are **NOT** part of the full-scale study sample of 25 clusters. The cluster's community context should be varied – one should be where most other program or child care options (beyond the available Head Start programs) are fairly comprehensive Head Start-like programs, one should be where there are program options which have some comprehensive program components, and one should be where there are few available options.

- Within each cluster we will vary the number of Head Start grantees/delegate agencies and/or centers that will be included in the field test to provide variation across auspices, number of unserved children, local service richness, grantee size, and urbanicity. We will combine these characteristics across 8 grantees where no two are alike on all characteristics. By selecting grantees with a diverse set of characteristics that are intentionally targeted we will be exploring the potential impediments to random assignment as well as testing study data collection procedures in a variety of settings.

Exhibit 5 presents our target design plan for the field study. It includes 24 centers across eight grantees. This configuration will provide the opportunity to test our data collection plans for the full study, and the ability of one site coordinator and measurement team to serve three grantees and nine centers. It will also provide the ability to test saturation issues in eight different contexts. Finally, varying the size of the grantee and the number of centers selected per grantee will allow for observations of the effect of burden on the ability of grantee/center staff to participate in the experiment. Every effort will be made to find grantees/delegate agencies that meet these combined characteristics. However, this exact design should be considered as our target goal rather than an absolute requirement. Input on site selection will be requested from Head Start Bureau and ACF evaluation staff.

Grantees/delegate agencies and centers will be selected and recruited for the field test during the **Winter of 2001**, and the family/child enrollment and random assignment processes will be implemented in the **Spring of 2001** with complete data collection beginning in **Fall, 2001**. We expect to sample about 600 children across the 24 centers. This will provide us an adequate number of cases to do some preliminary analyses and learn something about the differences in testing the treatment and control group children as well as the types of alternatives that families of control group children will choose in 24 different communities.

Exhibit 5: Target for Pilot Study Design

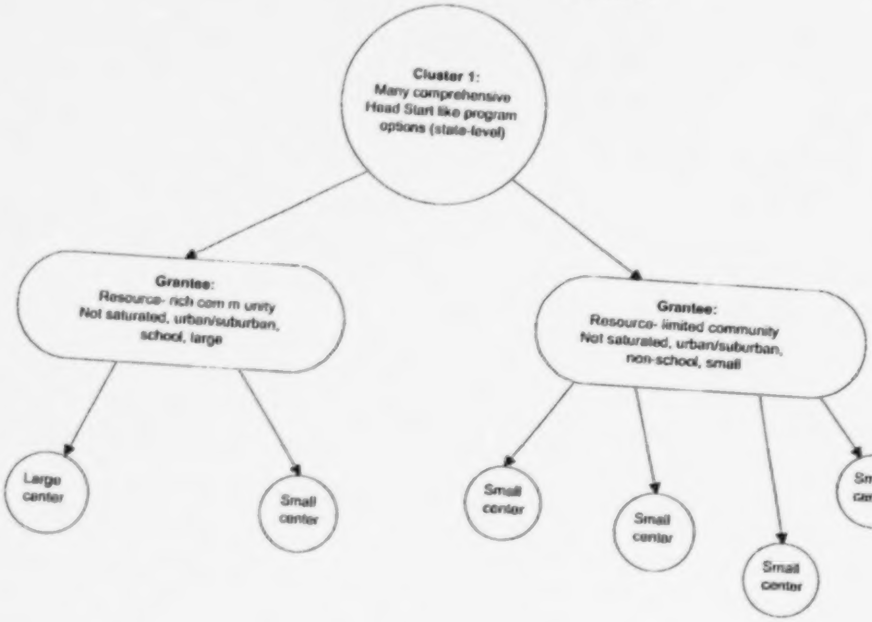

Cluster 1:
Many comprehensive Head Start like program options (state-level)
Grantee:
Resource- rich community
Not saturated, urban/suburban, school, large
Grantee:
Resource- limited community
Not saturated, urban/suburban, non-school, small
Large center
Small center
Small center
Small center
Small center
Small center

Exhibit 5: Continued

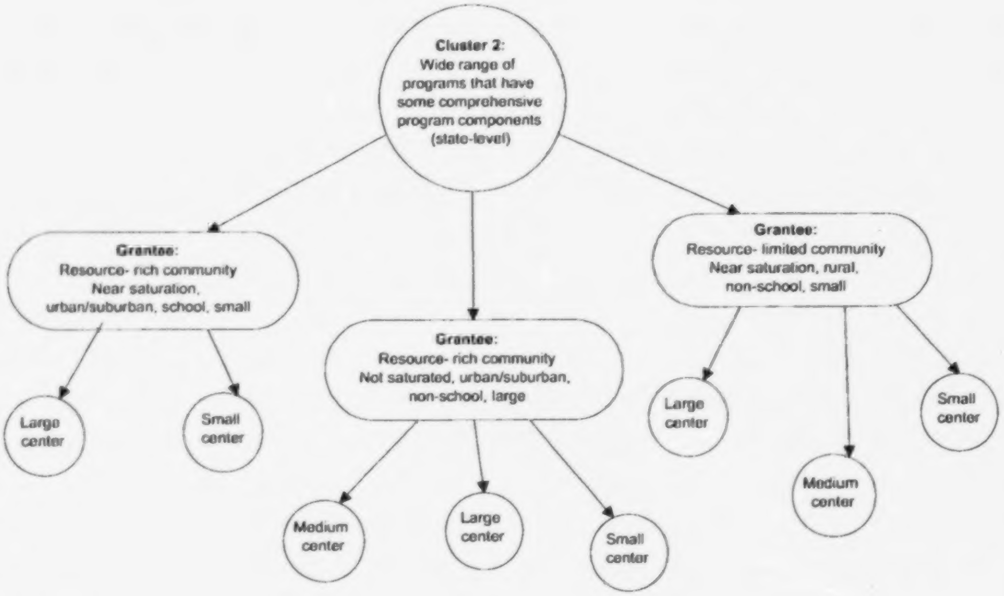
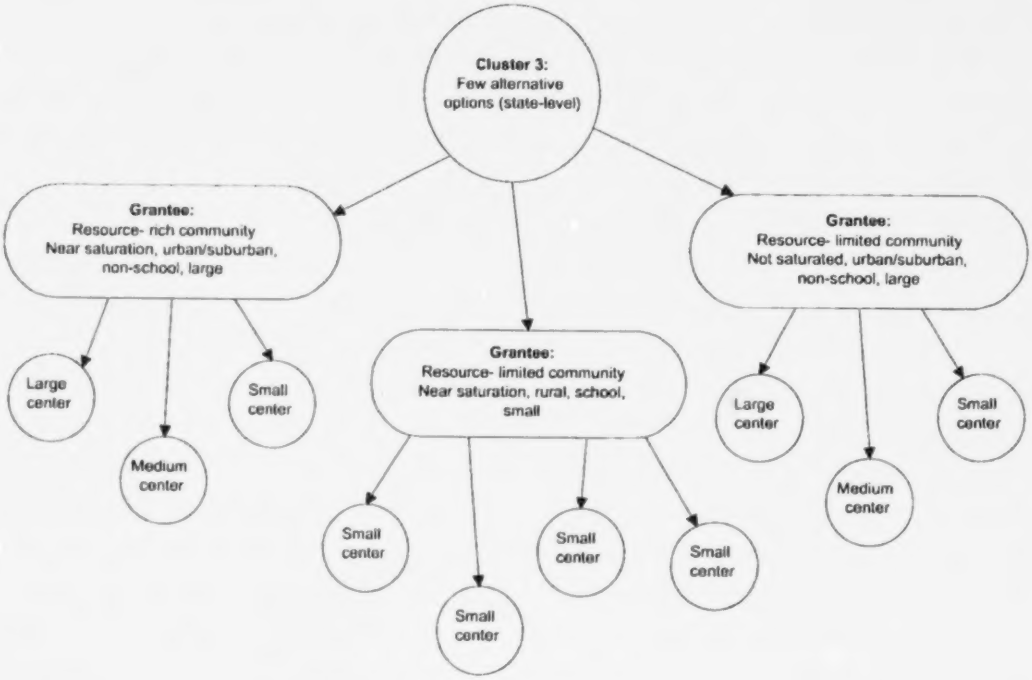

Cluster 2: Wide range of programs that have some comprehensive program components (state-level)
Grantee: Resource- rich community Near saturation, urban/suburban, school, small
Large center
Small center
Grantee: Resource- rich community Not saturated, urban/suburban, non-school, large
Medium center
Large center
Small center
Grantee: Resource- limited community Near saturation, rural, non-school, small
Large center
Medium center
Small center

Exhibit 5: Continued


Cluster 3: Few alternative options (state-level)
Grantee: Resource- rich community Near saturation, urban/suburban, non-school, large
Large center
Medium center
Small center
Grantee: Resource- limited community Near saturation, rural, school, small
Small center
Small center
Small center
Small center
Small center
Grantee: Resource- limited community Not saturated, urban/suburban, non-school, large
Large center
Medium center
Small center

4. Site Recruitment Strategy

The core of the proposed approach to recruiting grantees/delegate agencies for the study is to establish strong partnerships with the grantees/delegate agencies, actively address potential concerns, especially possible concerns of the Policy Councils, and work to make such issues as random assignment as acceptable as possible to both Head Start staff and participants. We have, therefore, proposed to engage our most senior and experienced project staff in this effort, drawing from a proven set of tools and strategies, as well as valuable information that will be derived from the work in the pilot study sites. These include the following:

- Early and ongoing communication and support provided from the Head Start Bureau and Regional Office staff.
- Gaining the support and endorsement of the National Head Start Association.
- Conducting sample selection and establishing contact with selected grantees early so there is time for problem solving, mutual learning, and the essential give-and-take needed to arrive at a design acceptable to all.
- Budgeting for ample advance on-site time to explain and motivate the study and move local program leaders into a shared belief in the study's importance and the essential nature of random assignment.
- Budgeting adequate resources to support the programs and thereby minimize the burden placed on any program for participating in the study, as well as reduce the likelihood of any disruptions to the quality of program service delivery.
- Softening the ethical objections to random assignment by allowing staff to provide information to control group families on alternative services in the community, providing **limited** exemptions from control group assignment for severely "at-risk" disadvantaged children meeting criteria specified in advance, and emphasizing the basic fairness of allocating Head Start's limited capacity among many deserving families by using a lottery method of selection.
- Removing the more onerous aspects of random assignment for line staff, including face-to-face notification of non-selection to Head Start and loss of influence over service targeting.
- Emphasizing the advantages of randomization from a program standpoint—its ability to show the value added of program services and point toward more effective service strategies for the future.
- Presenting the participation negotiation process as a give-and-take involving commitments and sacrifices from both sides, including "give" in the research design

where modification that does not compromise study reliability (e.g., reductions in the share of study subjects assigned to the control group).

- Holding down the number of control group cases sought from any given grantee, delegate agency, or Head Start center, so as to minimize the potential burden or disruption to normal program operations.

- Making resources available to grantees having problems meeting the requirements of random assignment—presentations to community groups, scripts for dealing with control group families, and training and/or financial support if expanded outreach is necessary.

- Enlisting peers as allies in the recruitment process—program administrators and policy makers, preferably from the Head Start community, with previous, positive experience participating in random assignment evaluations.

Immediately following approval of sample selection, recruitment teams will visit the GGCs and proceed to recruit Head Start grantees and gather information on the state and local context within which the grantees/delegate agencies reside. A site coordinator will play a critical role in recruitment, particularly after the recruitment team makes its first visit. It has been our experience that often times concerns about such issues as random assignment do not surface during the first meeting. It is important to have a local staff member on site to pay an immediate visit to program administrators to begin the process of establishing a working partnership relationship, review the scope of the study and to identify any potential concerns. The site coordinator will be able to alert the Recruitment Team if Head Start administrators are beginning to have "second thoughts." We will then be able to have a senior staff member talk become involved, as necessary.

Sample selection, establishing contact and developing a working partnership with all grantees/delegate agencies will be conducted early, at least 6 months before the earliest target date for the beginning of random assignment, and 12 months in advance wherever possible. It is anticipated that the recruitment process will entail ongoing contact with the sites via personal visits and telephone. Each recruitment effort must work through any potential concerns about participating in the study, develop individualized study plans with the grantees, and obtain information on the community context (as discussed below).

Local recruitment also includes gaining participation of providers of children in the control group. Once the sample of children is selected it will be incumbent upon the recruitment team to develop community plans for recruiting the various types of control group providers. Information obtained from the pilot study is expected to help with the development of a range of strategies and incentives used to recruit and retain the various types of non-Head Start providers. Similarly, the feasibility test will help refine or develop strategies to meet the Advisory Committee's recommendation to "collect the same or comparable information on children in Head Start and control group or comparison children (e.g., services received; quality and intensity of the intervention; and cost, descriptive, and contextual information.)"

A further key to successful site recruitment is to ensure program administrators that the use of random assignment will not impose too many burdens on potential participant families, nor generate dissatisfaction in this vital client population. The largest step toward this goal is the decision to conduct random assignment only in Head Start grantees/delegate agencies operating at near full capacity and where there is a pool of unserved families known to be interested in services. This ensures that the evaluation will assign children to the control group only where Head Start grantees/delegate agencies already have to turn down eligible applicants. Thus, there is no added burden on parents as a group. Moreover, the experience of other randomized studies suggests that when applicants must be turned away, acceptance of the lottery approach comes easily and almost universally. Some families even prefer it to a more complex, unseen process that can be viewed as capricious or even discriminatory.

With control families free to apply to other, non-Head Start sources of childcare and pre-kindergarten services, we anticipate few if any problems securing families' cooperation and maintaining their comfort and involvement in the study over its full course. This is strengthened by our proposal to set an early cut-off date for notifying families of the random assignment decision to allow them sufficient time to make other arrangements (see below). We also expect to be able to convince grantees that this will happen. Part of that assurance will come from involving agency staff in the focus groups of Head Start parents we have planned for the design phase of the project, an occasion to learn more about families' expectations, suggestions, and concerns regarding random assignment and refine procedures accordingly.

As previously discussed, a final strength of the proposed design that concerns grantee/delegate agency participation is the very small number of service exclusions required in generating the control group at any one grantee/delegate agency, relative to the overall scale of its operations, on average an increase of about seven percent. Compared with the demands of many other social experiments—particularly those that encountered difficulties convincing agencies to randomly assign applicants to a no treatment condition[44]—this is a very modest request.

Because the only Head Start grantees asked to consider these arrangements are those with more prospective participants in their communities than they can serve, it is hard to see why grantees at or above average size would object to excluding 1 in 15 applicants using a randomized lottery. In all likelihood, exclusions at that level are inevitable just in running the program in the normal fashion. Smaller grantees—where we would still be looking for 30 control group cases if possible—might well resist the initial suggestion to use random assignment. But even an extreme case, such as a grantee one-quarter the size of the typical grantee or delegate agency, would be assigning only 1 in 5 cases to the control sample and would need to identify only 27 percent more eligible children than currently served to make this possible. By using the many tools described above and our successful experience recruiting sites for other randomized studies—which made much greater demands on program operators—we believe we can hold to a handful the number of grantees lost to the study because of concerns about random assignment.

---

[44] For example, the National JTPA Study required participating agencies to raise intake by 50 percent rather than 7 percent and then excluded one of three applicants at random. Site recruitment went much better for the National Job Corps Experiment—in principal an equally difficult sales job but where only a very small number of control cases were required in relation to overall program size in any one site.

5. Random Assignment of Children

The first thing to recognize about the feasibility of implementing random assignment is the importance of collaboration between the research team and local Head Start program managers and staff—the evaluators must control the *designation* of treatment status for each child, while local program staff control its *actuation*. Thus, in practical terms, random assignment consists of three parts: (1) a statistical determination of treatment or control status for each potential Head Start child (the evaluator's job), (2) enrollment in, or exclusion from, Head Start in accordance with study status (the program operator's job) and (3) ongoing monitoring of the child's treatment or control status, as needed (both the program operator and evaluator's job).

For this process to work smoothly, program staff must understand the purposes of the evaluation and of randomization in particular, and staff must—at some level—endorse the study and its methods as worthwhile. This requires voluntary participation as any other basis for involvement will undermine working relationships between the partners and over time erode or disrupt research goals. With voluntary participation, program attrition—a grantee/delegate agency dropping its commitment to the study mid-way through the evaluation—should be rare.[45] Moreover, as shown in Exhibit 6 a feasible randomized design must be (1) acceptable to grantee/delegate agency staff and participants; (2) integrated into the normal enrollment process; and (3) useful to the broader policy and program constituency. The evaluator's role is to ensure that all these conditions are met within the unique circumstances of each local program. These are discussed below.

In general, grantees/delegate agencies receive most of their applications by early to mid-Summer then, following some cut-off date, make the bulk of their enrollment decisions at once, notifying those families that cannot be enrolled due to capacity limits (in many sites, these children are placed on a waiting list). Our random assignment design maintains this process to the point where children whose parents have applied for admission to Head Start are identified as eligible based on the individual targeting criteria of the local Head Start grantee.

To implement random assignment, we would ask grantees to select a larger number of cases at this step, still relying on their local criteria to select among eligible applicants. As discussed earlier, this number will be only slightly larger for the average-sized grantee or delegate agency (on, average about a 7 percent increase).[46] We would then obtain a list of these individuals and split them at random into three groups: 48 treatment group cases, 32 control group cases, and a residual group of non-study cases.[47]

---

[45] We do not expect any attrition from the random assignment process over the course of the study. Random assignment will take place at a single point in time following the experiment's application cut-off date of about June 30. Unlike many other experimental evaluations, no further random assignment support is required of participating grantee/delegate agencies.

[46] Since the number of desired control group members is fixed at about 32 per grantee/delegate agency, a somewhat larger expansion of the eligible list will be needed in small sites (e.g., a 27 percent expansion for a grantee one-fourth the average size).

[47] There will be 336 non-study children in the average-sized grantee or delegate agency.

Exhibit 6: Feasibility Requirements Of Random Assignment

Acceptability	■ Grantees see the value of the research. ■ Applications exceed capacity (or can be made to do so through slightly expanded outreach that does not fundamentally change the program or population served). ■ Ethical concerns about service "denial" can addressed by viewing random assignment as an even-handed way to allocate limited services across a broader applicant group (essentially, a "lottery") and/or offering families of control group members information on alternative service providers. ■ Consultation and protocols are provided by the evaluator to help program operators notify and take questions from cases assigned to the control group. ■ Staffing and funding burdens of assignment are effectively addressed
Feasibility	■ All applications flow through a central point where treatment and control determinations can be made. ■ Evaluator provides a secure (and "masked") mechanism for treatment/control determination that has the desired statistical properties. ■ Treatment/control determinations are made in real time, in keeping with the normal applicant screening and notification schedule of Head Start offices. ■ Monitoring systems can be put in place to ensure control group members are not served or later readmitted to the program.
Utility	■ Treatment/control status designations are upheld by program staff as enrollment takes place and services are delivered. ■ Participation "embargo" period for controls lasts long enough for important program impacts to emerge. ■ Non-Head Start services in community differ meaningfully from Head Start services. ■ Indirect influences of the intervention—those not differentiated between treatment and control group members (e.g., Head Start's potential community-wide impact on childcare quality from all sources)—do not create major program benefits.

BEST COPY AVAILABLE

Grantees/delegate agencies will then be asked to enroll the treatment group members and non-study cases for the Fall, and to not serve the children assigned to the control group, notifying them as they would any other non-admitted applicant.

Ideally, it would be preferable to conduct random assignment the day before the Head Start program begins in the Fall. That way we would be randomly assigning from the pool of parents and children who are ready to enter the program (i.e., this would also eliminate any dropouts between the time that applications are submitted and evaluated by the local program staff, and random assignment at the start of the program year).[48] But, such a strategy would, in our view, be very unfair to those parents whose children are assigned to the control group. That is, they would have to suddenly scramble to find alternative childcare arrangements for their child. Moreover, Head Start grantees/delegate agencies would see this as a problem for their parents, and we would, as a consequence, likely have a higher rate of program non-participation with the study.

The current plan is to set a cutoff date by which we would assemble a list of program applicants and use that list to select and randomly assign our study sample. Applications submitted after this date (including those admitted during the year to replace "drop outs") would go through the regular approval process but would **not** be eligible for inclusion in the experiment. We have arbitrarily set this cutoff point as mid-Summer. We chose this date to give parents a reasonable amount of time, after learning of their inclusion in the control group, to find alternative care options for their child. There is, however, nothing sacred about this particular date. Our actual plan is to work closely with each of the selected grantees/delegate agencies selected for the experiment and to see which arrangements and cutoff dates would work best within their established application process. Program-to-program variations in the cutoff date will not have any significant consequences for the experiment, as long as there is no bias introduced from the use of a cutoff date itself. As part of our planned pretest activities in all study sites, we will examine the extent to which the early and late applicants are different and the effect this might have on the representativeness of the study sample.

To strengthen the analysis of service types and Head Start impacts, we will also want to associate both treatment group and control group members with a specific Head Start center (i.e., to obtain measures of impact by center). In normal circumstances, no direct association exists between applicants who are *not* admitted and enrolled in Head Start and a particular program center, since grantees typically administer a single intake process for all centers. Program entrants are associated with specific centers, however, and may differ systematically in terms of program services, the social and economic status of Head Start families and the characteristics of their children, and the sources of childcare in the community. Hence, outcomes for Head Start children—and the outcomes that would have taken place for those children absent Head Start—are expected to emerge in distinctive ways at different centers.

---

[48] Early findings from the ACYF-sponsored ***Missing FACES*** project indicate that, on average, about 4% of admitted applicants never receive services, and about 14% leave during the program year, receiving only partial services. Under our plan, all children — regardless of how little, if any, services they receive — would be counted in the treatment group if that is where they were initially assigned at random. The "average" impact properly includes such partial services.

Associating children with centers can be accomplished in a number of ways, depending on local circumstances and the preferences of grantee staff. One option is to associate all Head Start applications taken prior to the cut off point with specific centers based on home address. The relatively small number of Head Start participants that apply after the cut off point, or enter the program after the school year has begun (e.g., to replace program dropouts), would not be included in random assignment or the study. Recognizing that grantees assign children to centers based on other factors such as parental preference and busing plans, these distinctions can also be applied to individual applicants where known, including those assigned to the control group.

A second, more refined way of associating control group children with individual centers would take account of the added role of capacity limits and "open slots" in making center assignments. To simulate this situation for both treatment and control groups, the children at the head of the eligibility list would not be assigned a treatment or control designation initially, but simply split in half at random at the application cut-off date. Grantee staff would then examine one-half the list and apportion its members to centers as though it constituted the service population for the Fall. The same exercise carried out for the second half of the list—again without knowledge of which half will actually enter the program—lines up both sets of children with specific centers. Following these designations, coded as special variables in the evaluation data base, a random "coin flip" by the evaluator would officially establish half the list as the true group of Head Start participants "pre-assigned" to individual centers—centers that can also be associated with counterpart members in the control group.

Keeping these issues in mind, our general approach to random assignment includes the following:

- Informing **all** applicants (at the time of application) to Head Start that slots are limited and that applications are due by a certain date at which point children will be randomly selected for inclusion into the program. A lottery approach will be used for selection into the study and subsequent random assignment to either a treatment or control group. Grantees will determine child eligibility and identify those children who will not be part of the study because of the high risk. As discussed earlier, these criteria will be carefully defined with the sites during the recruitment phase.

- On a predetermined **date** (or whatever is agreed upon with grantees), a list of eligible applicants will be supplied by the grantee staff and random assignment will be conducted by the study team.

- Site coordinators **will** monitor and verify the appropriate handling of sample members by the Head Start program and compare printed notification letters against a master list of treatment and control cases provided by the study team before letters go to the families.

- Letters will be sent to parents to let them know if their child is to be enrolled. Information about other available preschool or day-care grantees/delegate agencies will be provided to families who are not selected.

None of these procedures are "set in stone" at this point, and considerable flexibility will be needed to accommodate the needs and preferences of individual grantees and their delegate agencies and centers. Obviously, if any sampled grantee uses a different intake system than the characteristic one described above, a new random assignment process will be needed. Employment of on-site liaisons as key members of the evaluation team will add understanding and flexibility to our responses to these situations. Still, one point of consistency must be maintained wherever possible: the placement of random assignment in the intake flow prior to the start of services.

Informed Consent

All parents must be informed at the time of recruitment that a study is going to be conducted that might affect who receives Head Start, especially for control group children who would be "embargoed" from subsequent program participation. Therefore, information about the study and its potential effect on enrollment will be provided to ALL potential applicants. Parents need to understand and give their consent to the assignment process (i.e., before submitting an application that for some of them their children will not be eligible, for others there will not be enough slots, and those selected for the control group cannot be placed on a waiting list). It would be inappropriate to try and develop recruitment procedures that do not address these issues in a straightforward manner with grantees and families. Rather, it will be best to emphasize the points that random assignment will result in the same number of children being served by the program, that relatively few children will be assigned to the control group, and that families will be presented with information about other resources in the event that their child is not selected for enrollment in the Head Start program. A second level of written informed consent will be obtained from parents to participate in the full range of study activities after they have been selected into the study.

It also is recognized that the potential exists for the need to obtain a different level of informed consent from the parents of other children (those not in the study sample) in the various childcare settings and classrooms. Where this has arisen, Westat has addressed the situation successfully by preparing a letter to parents informing them of the study and its objectives and procedures. A passive consent approach is employed with parents notified of the dates and times that study activities are scheduled to occur and asked to reply only if they object to their child being present during and/or participating in these activities. In the event that a parent does object, their child is removed from the classroom and supervised in another classroom and/or setting as the particular situation permits.

Monitoring Random Assignment

A critical component to maintaining the integrity of a randomized study is consistent monitoring. Monitoring procedures will serve to track breakdowns while at the same time identify unanticipated problems. To accomplish this, a number of steps will be put into place, including an on-site coordinator, a system that tracks cases to their assigned treatment or control group, and a programmed random assignment methodology that will not allow double entry of a case.

The first step in monitoring is to set up a random assignment procedure that will be conducted completely by Westat staff and will not require the involvement of Head Start staff. To ensure that implementation of random assignment is carried out in accordance with the evaluation's needs, we will develop a manual of procedures, or guide, that documents how children will be selected, as well as data collection steps to distribute to local staff.

Part of the site coordinator's training, and documented in the site coordinator's manual, will be his or her responsibilities for systematically monitoring all the activities connected with rigorous implementation of random assignment. These activities will include, but not be limited to the following:

- Procedures will be established for enrolling children into the study. Only the site coordinator will be allowed to enroll children.

- Unique identifiers will be assigned for each case so that tracking information can be developed that will track each client through each phase of the evaluation.

- Checks will be made for other siblings who may have already participated in Head Start, as well the presence of younger siblings who may reach entry age during the course of the study (see the earlier discussion on this point).

- A check of enrollment and attendance records will be run within 2 weeks of the arrival of new participants in the Fall to identify any control group contamination (i.e., control group cases who "sneak in the back door" to obtain Head Start services and to identify any "no-shows").

- During the first year, site coordinators will meet with each Head Start site liaison on a monthly basis to review Head Start enrollment and attendance records to cross-check children enrolled in the control group. We will check at the grantee level. If there is more than one grantee per GGC, we will also check attendance records for these grantees. For the 3-year-old population, this will also be implemented in the second year.

- Before randomization, site coordinators will talk with Head Start personnel on an ongoing basis to answer questions about the study and help reassure them about concerns. Building these relationships from the beginning helps to encourage open communication and to identify problem situations.

Experience tells us it will be difficult to achieve an "ideal" experimental design. We cannot control for everything. However, by building in safeguards throughout the process, we can identify breakdowns early and develop new procedures as problems occur. Furthermore, having identified such problems, we can attempt to counter their influence in data analysis. For example, we can test the sensitivity of randomized study findings to such breakdowns by making alternative assumptions about how such cases would have behaved had their random assignment not been subverted.

Use of Incentives

An important part of gaining study participation is the use of incentives. For the Head Start Impact Study, we are proposing differential incentive payments in the form of gift certificates for various respondents. The site coordinator will be responsible for identifying, in consultation with the on-site liaison, the appropriate vendor (e.g., a food store for parents or a book store or toy store for teachers and classrooms) and purchasing the certificates in a variety of denominations to allow combinations of certificates to be provided for different levels of incentives.

During the preschool years, all Head Start classrooms participating in the study, and for the control group, all other care providers who will be observed for quality assessment will, ***pending approval by the Office of Management and Budget,*** receive $25 gift certificates for a toy or book store in their area in appreciation of allowing us to visit and for their overall cooperation. In addition, pending approval, all teachers and care providers will be given $10 gift certificates for the completion of their questionnaires and an additional $5 for the completion of each individual student assessment. While some teachers will receive a higher award than others, fairness will be achieved by providing teachers with awards based directly on and in proportion to the level of effort required. Westat intends to award parents and primary caregivers gift certificates to a local supermarket, an incentive we believe to be both practical and attractive to this group of respondents. All parents and primary caregivers will (pending approval) be awarded $20 gift certificates for completing their interviews. If the parent assists our efforts by contacting the child's care provider and securing the care provider's participation before project staff contacts them, an additional $15 gift certificate will be provided to the parent, a procedure that has proven successful.

During the kindergarten and first grade years, all teachers will (pending approval) be given $10 gift certificates for completion of their self-administered questionnaires and an additional $5 for each student assessment. Parents will again earn $20 for completing their interviews. Finally, we intend to provide small gifts for the students such as stickers or pencils. We have used these incentives effectively on other school studies to enhance student interest, increase motivation, and ensure high rates of participation.

The site coordinator will distribute the incentives in person or by mail as appropriate. He or she will be responsible for maintaining complete documentation of all awards earned and disbursed.

As noted above, the proposed dollar amounts for incentives are subject to review and approval by OMB.

6. Data Collection

6.1 Overview

The Head Start Impact Study is a multi-faceted and complex longitudinal research effort that requires a comprehensive data collection plan that structures and unifies the various requirements into a well-defined plan of action. The key to successfully implementing a challenging longitudinal research effort is a data collection plan that emphasizes flexibility within a thoughtful structure. Westat's Operations Director will be responsible for implementing the plan and coordinating the efforts of key individuals and highly experienced, skilled teams to complete all recruitment, random assignment, data collection, and monitoring/quality control tasks in an efficient, organized and timely manner. To implement this approach, we created two major roles:

- **Site Coordinators.** Selected from Westat's cadre of experienced field supervisors, staff will be assigned to each GGC, immediately after sample selection and for the duration of the study. Their role will be a pivotal one serving as the primary local contact; assisting the recruitment team with securing participation of grantees; facilitating random assignment; enlisting cooperation and maintaining participation of respondents, including parents, children, providers, teachers and administrators; coordinating all data collection activities in the GGC; tracking study participants; managing field staff; and ensuring QC. They will report to the central office operations director and her staff. We expect to hire one site coordinator for each GGC, for a total of 25 site coordinators.

- **Measurement Team.** Consisting of Westat field interviewers, under the management of a site coordinator, these teams will be responsible for scheduled data collection activities for each wave, including conducting in-person parent interviews, child assessments, in-person staff interviews, and assessments of program quality.

We also plan to ask each grantee/delegate agency to identify a staff person who as an **on-site liaison (supported at least in part by funds from the evaluation contract)** will serve as the main point of contact with the Head Start program. The on-site liaison will be invaluable in helping our site coordinators to establish rapport, obtain the cooperation of other staff, secure the trust and cooperation of parents and the community, and facilitate random assignment and data collection.

Data collection will focus on the full range of comprehensive services and integrated program elements for children and their families that form the cornerstone of the Head Start program and contribute to the child's readiness for school. We will collect comparable data on two cohorts of children (a 3-year-old cohort and a 4-year-old cohort) and their families who will be randomly assigned to either a treatment group (to enroll in Head Start) or a control group (that will not enroll in Head Start, but will be permitted to enroll in other available services selected by their parents or be cared for at home). From each of 75 grantees/delegate agencies, we plan to select a sample of approximately 48 children per grantee/delegate agency for the Head Start treatment group. To these about 32 children will be added who will be part of the control group. All children must not have been previously enrolled in Head Start. The selection and random assignment of children will occur during the **Spring/Summer of 2002**.

Data collection will begin in **Fall 2002** and extend through **Spring 2006** with waves in the Fall and Spring of the Head Start year(s) and in the Spring of the kindergarten and first grade years. During each wave, we will administer assessments to children. In addition, to obtain valuable information that only the parent or the child's primary caregiver can provide and to keep parents/primary caregivers actively involved, we will conduct face-to-face interviews with them, twice a year in Fall and Spring, for each year of the study. Once a year during the Spring waves, we will interview program administrative staff, survey teachers/other care providers, and where necessary, collect data from administrative records. Each Spring, we also will conduct observations and assessments of the quality of both Head Start and other care settings. The field period will vary in length up to two months, depending upon the scheduled activities. The length of time that we will be at a given site will vary with the number of children to be assessed and the activities for the particular wave of data collection. Exhibit 7 provides a summary of the data collection schedule.

We anticipate that some grantees/delegate agencies may be excluded from random assignment due to unwillingness to participate or saturation. It will be important to check for non response bias due to grantee exclusion. Our plan is to obtain data on up to five non-participating grantees/delegate agencies, 15 associated centers and an average of six Head Start children per center. Both three-and four-year old children will be included, and we will collect the exact same data from these agencies, families, and children as for the full study. As our main concern is to determine the bias associated with non-participating sites, families, and children, we will only collect Fall and Spring baseline data. We will also use FACES data to augment our understanding of non-participating sites (see Appendix B).

During the **Fall of 2001** and **Spring of 2002**, we will conduct a **field test** of all procedures including those for selecting and randomly assigning children and families to treatment and control groups, notifying parents/primary caregivers, obtaining informed consent, identifying other childcare settings and securing cooperation from the various respondents. We will test all data collection instruments and conduct focus groups of parents, Head Start program and other childcare providers to evaluate such issues as comprehensiveness and sensitivity of proposed questions.

Each phase of data collection will take into account the variations in Head Start settings and experiences of the treatment group children as well as the childcare settings and experiences of the control group children. It will be incumbent upon the recruitment team and site coordinators to identify the variations in each of their assigned communities so that adaptations can be developed as necessary.

Exhibit 7: Data Collection Schedule

School Year Preschool year/grade for C1 Preschool year/grade for C2		2002-2003 3-year-old Preschool 4-year-old Preschool		2003-2004 4-year-old Preschool Kindergarten		2004-2005 Kindergarten Grade 1		2005-2006 Grade 1	
Data source	Cohort	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Children	C-1	✓	✓	✓	✓		✓		✓
	C-2	✓	✓		✓		✓		
Primary Caregivers	C-1	✓	✓	✓	✓	✓	✓	✓	✓
	C-2	✓	✓	✓	✓	✓	✓		
Administrative Records	C-1		✓		✓		✓		✓
	C-2		✓		✓		✓		
Program Staff/Other Care Providers and Elementary School Teacher	C-1		✓		✓		✓		✓
	C-2		✓		✓		✓		
Quality of Care Settings	C-1		✓		✓		✓		✓
	C-2		✓		✓		✓		

KEY
Cohort 1 (C1) – 3-year-old cohort
Cohort 2 (C2) – 4- and 5-year old cohort

6.2. Data Collection Strategies

Planned Data Collection Activities

After selecting the 25 GGCs, we will contact both regional offices and grantees/delegate agencies by letter and telephone to determine the programs' eligibility for the study. Using PIR data and a variety of secondary sources, we will verify whether the grantee/delegate agency meets the rules for inclusion in the study. Subsequently, recruitment teams composed of the most senior staff integrated across all four companies will visit each grantee cluster, to foster partnerships and enlist the cooperation of the grantees/delegate agencies as well as to verify and obtain additional information on programs, their centers, and the communities within which they reside. We feel it is crucial to gain an understanding of the state and local context within which the Head Start programs operate. Consequently, we will conduct a case study effort using multiple primary and secondary sources to investigate and describe, among other variables, the types of services available, the extent of coverage, and the patterns of childcare/early education usage in each selected GGC.

Data Collection Sources

Parents/Primary Caregivers. Face-to-face interviews, no more than 1 hour in length, with parents/primary caregivers will be conducted twice a year each year the child is in the study to obtain information at more frequent intervals and keep parents actively involved. We propose to interview the primary caregiver of the child—in most cases, the mother, if she is present in the home, or the biological father, or, failing that, a knowledgeable substitute parent. We expect to conduct a limited interview in the Fall and an expanded version each Spring with the one exception of the baseline Fall interview that will be more comprehensive in scope and similar in length to the Spring instruments. Parent interviews will be conducted in the languages of the respondents, and we will have a Spanish version of the interview. For all other language possibilities, we will rely on bilingual field interviewers or, if necessary, we will enlist the aid of an interpreter from the Head Start program or from the neighborhood to translate and conduct the interview.

Children. Children will be assessed individually in the Fall and Spring of the Head Start year(s) and each Spring of the kindergarten and grade one years. Assessments will vary over time consistent with the child's age and grade. We are considering drawing on our experience with FACES and using a battery of some of the same child measures employed in this study. We will convene a panel of experts to review these instruments and make recommendations for improvements.

We expect the child assessment to take 30-40 minutes per child. Finding the appropriate place to conduct assessments will require flexibility on the part of our measurement team. For example, they may need to arrange, for both the experimental group (including those in Home-based Head Start options) and the control group, that assessments be conducted in a variety of settings including homes or at central locations. They also may need to arrange for transportation. (See the next section on Measures for a description of the planned instruments.)

Program Staff, Teachers/Other Childcare Providers. Program staff, teachers/other childcare providers will be surveyed each Spring. Each staff and teacher/other childcare provider survey will require no more than 25-30 minutes in length. We anticipate conducting interviews with the Head Start center director and the family services coordinator (or the individual who plays that role in the center), and with comparable staff as found in other care settings. With the child's teachers during the preschool and elementary school years, in light of the costs associated with conducting interviews with 3,000 to 4,000 teachers, we will conduct a paper and pencil survey rather than face-to-face interviews.

Since children will be exposed to a wide variety of types of care settings selected by their parents in addition to or in lieu of Head Start, it is necessary to arrive at a means of carefully targeting and limiting the number of care settings at which to collect staff, care provider and teacher data. We, therefore, will limit the number of care settings to a maximum of one per child and will select the program where the child is in care the greatest number of hours between 9:00 a.m. and 3:00 p.m., Monday through Friday. For children in the treatment group enrolled in Head Start, the Head Start program will be the chosen provider. While we will not directly survey staff at more than one childcare provider per child, we will obtain information from the parent/primary caregivers' interviews regarding the additional care settings the child experiences.

Teachers and childcare providers also will be asked to complete rating scales on each of the sampled children concerning their development and behavior, a task that will require a few minutes per child. Teachers and childcare providers will receive gift certificates as incentives for completing the interviews, questionnaires, and rating scales.

Assessment of Quality of Care Settings. Direct observations of the quality of programs or care settings for the Head Start and control groups will be conducted annually in Spring during the preschool and elementary years. We will limit our assessment of quality of care to the same setting where we survey staff, teachers and other care providers, as described above. For Head Start Home-based options, we expect to conduct observations both during socialization experiences as well as at home. We will include questions on the parent/primary caregivers' interview to elicit information from these respondents' perspectives on quality of additional care settings the child is in.

Administrative/School Records. While we expect to find formal records at Head Start programs and in schools, we anticipate that comparable data will not be maintained in all types of care settings, particularly less formal settings where a substantial proportion of the control group is apt to be cared for. We are cognizant of this issue and will include questions, for example, on the parent instrument to provide another source of data that we may not find in written record form. We will make use of computerized records and reports where available and/or abstract children's records each Spring to gain information on child attendance as well as useful tracking information.

How Will We Encourage Participation?

Head Start Programs/Other Care Providers. It is impossible to overstate the importance and challenge of recruiting sampled grantees/delegate agencies for participation in the evaluation. We are currently engaged in a variety of efforts to build awareness of the impact study and encourage interest in participation and cooperation. We are working closely with the National Head Start Association and are contemplating disseminating information regarding the study through their video training sessions. We attended the National Head Start Child Development Institute where we participated in the poster session and distributed over one thousand fact sheets describing the study's goals and objectives. We plan to attend additional conferences in the near future including Head Start regional, parent, and state director conferences.

As noted earlier, immediately after approval of sample selection, Westat will contact the selected Head Start grantee and delegate agency directors by letter. Following the mailing, each program director will be contacted by telephone by a recruitment team leader who will schedule the Recruitment Team's visit to the program to discuss arrangements for participation in the study. This visit is a key element in establishing trust and rapport while emphasizing the importance of the centers' cooperation and participation. In addition, the site coordinator will play a critical role in recruitment, particularly after the recruitment team makes its first visit. It is important to have a local staff member on site who can follow up and pay an immediate visit to program administrators to review the scope of the study and to identify any concerns about random assignment. By providing financial incentives, technical assistance and training for project staff, and emphasizing the benefits of participation to selected grantees/delegate agencies, we feel confident in our ability to persuade most sampled agencies to participate. Local recruitment also includes gaining participation of providers of children in the control group. Once the sample of children is selected it will be incumbent upon the recruitment team to develop community plans and strategies for recruiting control group providers.

Parents/Primary Caregivers. Planning to gain parent cooperation must focus from the start on both short and long-term participation in the survey. The first step in ensuring participation will be for the site coordinator to provide the program with packets of materials to distribute to parents upon enrollment. The packets will include an information sheet to inform parents of key elements and benefits of the study and enlist their participation in the study. We also plan to produce a video to explain the elements and benefits of the study including the process of random assignment.

Gaining participation of families who do not receive Head Start services will need to be handled even more judiciously. Our incentive program will be stressed and will be an important tool to help gain participation of all families. Rather than developing recruitment procedures that try to hide issues surrounding random assignment, we believe it will be better to emphasize the points that random assignment will result in the same number of children being served, that relatively few children will be assigned to the control group, and that families will be presented with information about other resources.

We will obtain complete written informed consent from parents after children and families have been selected into the study. Telephone follow-ups will be conducted to actively recruit parents and their children selected for the study. Offering incentives will be an important tool in securing parents' consent along with providing assurances of confidentiality and stressing the importance of having their voices heard. We will provide a practical incentive that can benefit the entire family, a gift certificate for food from the local supermarket.

Other Care Providers. Parents and primary caregivers have a key ongoing role in identifying the child's care providers and granting permission for us to contact them. Moreover, without some form of introduction and authorization from the parent, most contacts will be unwilling to provide information. In general, we have proposed several features designed to enhance the cooperation of the childcare providers including: collecting during the parent interview appropriate locating information for the provider; asking about any special arrangements that may be needed for contacting the provider; and developing a letter to the childcare provider to be signed by the child's parent that explains that the parent and child are participating in an important study, encourages the provider to participate, and authorizes the provider to supply requested information about the child. We will award a gift certificate to the parent as an incentive to make the first contact with the provider, and the provider, in turn, will receive a gift certificate for his/her participation.

We anticipate that other childcare providers will include parents, relatives, day care homes, other day care centers, and other pre-K programs. Although there will be similar concerns among these providers, materials and procedures will be tailored to individual circumstances.

Elementary School Teachers and Personnel. Our plan for gaining the cooperation and adequate participation of elementary school teachers and personnel for data collection is to begin early and to provide an advance publicity packet to districts and schools likely to be the recipients of children participating in the study to inform them of the study, the critical nature of the information schools and their teachers will provide, reasons to participate in the study and the incentives that respondents will receive. We will follow this up with telephone calls and/or in person visits by a Recruitment Team member and/or the site coordinator to answer any questions that administrators have regarding study participation and requirements. We will make every attempt to determine early any requirements that a particular district may have with regard to the conduct of research in their schools and to respond promptly

How Will We Maintain Contact With Families?

Ongoing tracking in longitudinal studies is critical to maintaining high response rates. We will follow all children selected into the treatment group even if they do not enter the Head Start program in the Fall and will track all treatment and control children even if they move out of the study area. We propose to follow and collect data from all movers within a GGC and to follow a 10 percent subsample of movers to faraway locations.

We will track and locate children and parents employing a variety of approaches that we have used successfully on other longitudinal studies with similar populations. We will obtain tracking information as early as possible in the study enrollment process and will update it constantly and at least every six months during each wave of parent interviews. All information will be entered into a database to manage efficiently all tracking activities. This database will contain a history of contact and locator information for each child included in the sample over the course of the study. As time goes on, some parents will refuse to participate and data collection staff will be trained to convert as many of these refusals as possible. The use of incentives will be an effective means for building and sustaining interest and participation.

7. Measures

7.1. Overview

A wide variety of data sources and measures will be used in this study to (1) assess the difference Head Start makes in the development of the nation's low-income children, and (2) identify the conditions under which Head Start works best and for which children. The measures will include:

- ***Child and family measures*** including a child assessment, a teacher's/'childcare provider's report on the child's approaches to learning and behavior, and a parent/primary caregiver interview focusing on their assessment of the child's readiness to learn and social competence as well as information on parenting skills and the available community services.

- ***Program measures*** including classroom/childcare observations focusing on dimensions of classroom/childcare quality such as personal care, furnishings, language and reasoning activities, gross and fine motor activities, creative activities, social activities, and provisions for adults and/or teachers; and center director interviews, family service worker surveys, education coordinator surveys, and teacher/childcare provider surveys that focus on demographic characteristics and the perceived learning environment.

- ***Contextual measures*** including secondary data sources that will provide and confirm contextual and capacity information. The contextual variables will assist in understanding childcare at the grantee level including the state commitment to childcare options and the quality and quantity of other available options.

We propose relying primarily on the battery of instruments used in the Head Start Family and Child Experiences Survey (FACES) as well as those instruments that are applicable from ECLS-K. The Abbott Early Childhood Education Study in New Jersey is currently using the FACES instruments. The study team does not intend to carry out a major redesign of the FACES battery but we intend to convene a panel of experts to review the existing FACES instruments and to make recommendations for improving some of these instruments, as may be necessary. We also intend to review the enhancements that the FACES team has recommended or is trying to implement in FACES 2000. These studies have pulled together instruments from the top researchers in the country, designed to cover the important domains of social competence as well as the environmental factors that influence child development. The following sections briefly describe the various measures proposed for this study.

7.2. Planned Measures

Child and Family Measures

We propose to use the FACES child assessment battery for this evaluation through kindergarten and augmented, when necessary, for the first grade. The FACES child assessment consists of a series of tasks designed to appraise the children's cognitive and perceptual-motor development in areas such as word knowledge, letter recognition, and copying of designs and letters, tasks shown to be predictive of later school achievement, especially later reading and oral language proficiency. The assessment also examines the child's social and communicative competence (i.e., telling facts about self and family to another person). The following measures are included in the FACES child assessment battery: (1) the Peabody Picture Vocabulary Test, Third Edition (PPVT-III); (2) Letter-Word Identification, Applied Problems, and Dictation tasks from the Woodcock-Johnson Psycho-Educational Battery—Revised (WJ-R); (3) Story and Print Concepts task; (4) Draw-A-Design subtest of the McCarthy Scales of Children's Abilities; (5) Phonemic Analysis subtest of the Test of Language Development-3 (TOLD-3); (6) Color Naming and Counting task (developed specifically for FACES); and (7) Social Awareness items from the CAP Early Childhood Diagnostic Instrument.

The Child Assessment battery is available in both English and Spanish. In assessing children from Spanish speaking families, the study team is considering the use of the Pre-LAS (i.e., Pre Language Assessment Screener) as a screening tool for determining a child's dominant language. We also intend to consult with experts in the assessment of non-English speaking children to review the existing instruments and if necessary, identify additional screening tools and alternative assessment measures that will provide comparability with the measures used for the English speaking children. The procedure used in FACES 2000 is to assess the children from Spanish speaking families on two components of the assessment in both English and Spanish. These components are the Woodcock-Johnson Letter-Word Identification task and the PPVT-III. In the Fall of the Head Start year, children are given the Spanish version of the FACES battery and the English version of the two duplicated scales (i.e., Woodcock-Johnson Letter-Word Identification and the PPVT-III). In the Spring, the children are given the English version of the full battery, plus the two duplicated scales in Spanish. This should enable the study team to track growth (or decay) in language proficiency in both English and Spanish.

Given the focus of the Head Start Impact Study, the parent interview will include information in the following areas: (1) parental beliefs and attitudes towards their child's learning, and parental participation in and satisfaction with the program; (2) family household and demographic information including parent-child relationships and the quality of the child's home life; and (3) parent ratings of their child's behavior problems, social skills, and competencies. The parent interview in the FACES 2000 includes a subset of 13 items from the lengthier Child Rearing Practices Report (CRPR). These items address key areas that may be affected by parental exposure to Head Start, such as attitudes about how parental authority is conveyed and encouragement of the child's exploration and independence. Parents will be asked about services they receive and the help they receive in coordinating the services. Other topics will include the child's transition from preschool to kindergarten and any information or services the family received to assist with this transition. A number of items in the parent survey are drawn from the National Household Education Survey. This provides a national point of comparison for questions related to parent perceptions of the kindergarten and first grade programs, and the child's experiences in progressing through school.

In addition to the interviews, the parents or the primary caregivers may be asked to complete the Child Rearing Practices Report (CRPR), a 91-item questionnaire that assesses parents' attitudes, values, behavior and goals. The CRPR covers four general domains: (1) how positive and negative emotions are expressed, handled, and regulated; (2) how parental authority is conveyed, and the specific forms of discipline that are used; (3) the parent's ideal and goals with respect to the child's accomplishments and aspirations; and (4) the parent's values concerning the child's development of autonomy, independence and self-identity. Since the issue of overburden is a concern, during the pilot phase, we will decide whether the scale should be used in its entirety or whether the abbreviated FACES version should be used.

Teachers and childcare providers will be asked to rate each child participating in the study using the teacher's/childcare provider's child report form. Information is collected in the following areas: social skills, classroom conduct, problem solving and initiative, social relationships, creative representations, music and movement, and language and mathematics. The current FACES instrument will be reviewed and modified if necessary for use with the childcare providers.

Program-level Measures

Measures of classroom processes proposed for this study include the Early Childhood Environment Rating Scale (revised) (ECERS-R), supplemented by three subscales of the Assessment Profile for classrooms, as well the Arnett Scale of Teacher/Childcare Provider Behavior. This combination of measures has been used successfully by Westat in the FACES study. The ECERS-R provides a global rating of classroom quality based on structural features of the classroom including personal care, furnishings, language and reasoning activities, gross and fine motor activities, social activities, and provisions for adults and teachers. The Assessment Profile, a structured observation guide designed to assist in self-assessment to improve the quality of early childhood programs, measures important characteristics that are not easily captured by the ECERS-R. We propose the use of three subscales from the Assessment Profile: (1) Scheduling, (2) Learning Environment, and (3) Individualizing. The Scheduling scale requires the observer to make ratings based on the posted classroom schedule and it also measures the degree to which the teacher provides for small-group or individualized activities during the classroom day, and whether there is a mix of indoor and outdoor activities, and a mix of quiet and active activities. The Learning Environment scale is most crucial for identifying whether materials for child use in the classroom are both available and accessible to the child during free play. The Individualizing scale focuses on maintaining developmental portfolios on children and what the teacher and program do to track children's progress. Finally, the Arnett Scale of Teacher/Childcare Provider Behavior is a rating scale consisting of 26 items organized under five areas: sensitivity, punitiveness, detachment, permissiveness, and promotion of independence.

Although we may not be able to use all the scales in less formal settings, we do propose using them for all non-Head Start centers. To maintain comparability, some scales used in the more formal settings may be appropriate, while other scales may be modified for use in the less formal settings. If additional measures are needed for the less formal settings we will explore using the Family Day Care Rating Scale (FDCRS), which was developed by the same people who developed ECERS-R. It has been used in the NICHD Early Child Care Study. We will also examine using the newly developed Observational Record of the Caregiving Environment (ORCE), a conglomeration of a number of other scales.

At the kindergarten and first grade level we will implement a more limited approach to measuring quality. We will rely on information from secondary sources to track a school's record with respect to such issues as attendance, disciplinary issues, immunizations of children, average test score, number of children receiving school lunch, and teacher/student ratio.

We also expect to gather specific information about the child's experiences and development from the perspective of the teacher and/or childcare provider. Surveys will include questions to obtain biographical information including education and years of experience, inquiries regarding program elements, quality of management, and belief scales to assess staff attitudes on working with and teaching children. Items on literacy promoting activities, parallel to questions used in the ECLS-K, are included in both the teacher and center director surveys. Use of these items provides a national sample benchmark for the measures. During the kindergarten year, the teacher survey will obtain information about the kindergarten program, provisions that were made for the child's transition to kindergarten, and whether the teacher obtained any information from the Head Start program or alternative care provider about the child's development status or special needs.

The center director interview will provide additional information on the operation and quality of the program. Issues to be addressed in the interview will include: staffing and recruitment, teacher education initiatives and staff training, parent involvement, waiting lists and program expansion, curriculum, classroom activities and assessment, home visits, kindergarten transition, and demographic information about the director. Other sources of information on the operation and quality of the program are the FACES 2000 family services worker survey and the education coordinator survey. Issues addressed in the family services worker survey include services provided to parents, how family needs are assessed, most common family problems, and obstacles encountered in coordinating services while the education coordinator survey addresses issues such as program organization, staff education and training, educational philosophy, and curriculum and classroom activities.

Contextual Measures

Regional offices will be contacted to determine whether grantees are eligible for inclusion in the study. Specifically, we will determine that: (1) the grantee is not new (i.e., in operation less than two years); (2) the grantee is in compliance under Head Start performance standards (i.e., the grantee is not operating under a Quality Improvement Plan or formally designated as high risk status); and (3) the grantee is not operating substantially under-capacity or under enrollment. We also will ask the regional office about an "at risk" category or some similar identification, short of the term "deficient" that alerts the grantee that they have some problems, and if they are aware of any issues with a grantee that would make participation in the study problematic. We will try to verify this information through multiple sources of information. PIR data will be used extensively and while these data cannot be used solely to identify under-capacity and saturated sites, they will aid us in gaining an understanding of and pinpointing those sites that will likely require more attention and follow-up to verify their status.

We also will collect information from the grantees focusing on recruitment, selection, and enrollment. This information will be used to augment and verify data obtained from other sources including the PIRs and the regional offices. Specifically, we will ask questions about: (1) the number of funded slots; (2) enrollment patterns over the past 2-3 years; (3) filling funded slots when a child leaves the program; (4) waiting lists; (5) number of over-income families on the waiting list; (6) number of over-income children currently enrolled; (7) status of full-day services; (8) competition from other childcare centers or pre-K programs; (9) number of additional families that could be served with unlimited resources and sufficient funding; and (10) ability of the grantee to recruit 15-20 more Head Start eligible children than presently recruited.

In order to understand the service environment and options that are available to control group children, it is important to understand the communities in which the sampled Head Start programs operate. This effort will include the collection of secondary data at the national, state, and local levels, and the collection of primary data at the state and local levels. Possible secondary data sources include the following:

- Urban Institute's Assessing New Federalism (ANF);
- Administrative data from the PIR on Head Start enrollments, participant characteristics, and grantees;
- Additional data from the Child Care Bureau on childcare subsidies and programs;
- Data on state pre-kindergarten/Head Start initiatives from the Children's Defense Fund's *Seeds of Success*;
- Information about the extent of comprehensive state initiatives focusing on preschoolers from the National Center for Children in Poverty's *Map and Track*;
- Data currently being collected on state and local policies and investments by APHSA, DCF, NCSL, and other national organizations; and
- Ellsworth/Child Trends information on county level Head Start eligible and Head Start served children.

At the state and community levels, we will collect primary data to supplement and deepen the information available from secondary sources. From state level administrators, we will solicit information about the range of programs available and the availability of program statistics, especially those broken down at the substate or local levels. We will look for: (1) information about the childcare market, including licensed slots in childcare centers, licensed slots in family childcare homes, as well as to find out which facilities are not covered in the program statistics; (2) enrollment patterns in state childcare/early childhood programs, such as state pre-K programs, utilization patterns of state and Federal childcare subsidy programs; (3) key childcare/early childhood policies and initiatives that can shape the childcare market and context, such as collaborative initiatives between Head Start and pre-K or childcare, wraparound initiatives, etc.; (4) enrollment criteria for programs in terms of income eligibility and age range of children; and (5) how states have prioritized the distribution of subsidies among welfare families and other low-income families.

A key component in understanding the communities in which the Head Start programs operate will be information collected from key local informants. The information collected at the local level will focus on the following: (1) how the local childcare market works in the community; (2) the types of childcare and early childhood education programs available to low-income families in the community; (3) the supply of slots across various program types; (4) the hours the slots are available; (5) the out-of-pocket costs of slots; and (6) the criteria and mechanisms used to place children in them.

Summary

In summary, several tasks must be accomplished in order to complete the measures for the study. The tasks include the following:

- Convene a group of experts to review the existing FACES instruments and provide suggestions for improving the instruments.
- Convene a group of experts on the assessment of non-English speaking children to review the existing measures and if necessary, identify alternate assessment measures.
- Convene a group of experts for guidance on the development of telephone and site visit protocols for collecting contextual information.
- Develop telephone protocols for determining eligibility for inclusion in the study and for collecting information on recruitment, selection, and enrollment.
- Identify secondary data sources for determining the service environment and childcare options that are available in the selected communities.
- Develop site visit protocols for collecting information on how the local childcare market works in the community.

8. Analysis Plans

As noted above, there are two over-arching study goals—to estimate the national impact of Head Start participation on a variety of child- and family-level outcomes, and to assess the extent to which those impacts vary as a function of individual- and program-level characteristics. The plan is to define the treatment and control group, as discussed earlier, and to randomly assign children to one or the other alternative. The national impact estimate (i.e., the first goal of the study), derived from this experimental design, will then answer the question of the effect of the federally-managed Head Start program on eligible children. Under the second goal of this study it will be possible to examine how program impacts vary along a variety of dimensions, including the extent to which there is wide availability of Head Start-like programs in the community. Moreover, as discussed below, it also will be possible to use quasi-experimental techniques that capitalize on the randomized design to estimate the contribution of the Head Start service model to child outcomes compared to less intensive services even in places where control group members in some instances received various forms of comprehensive services or other non-federally funded Head Start-like services.

The remainder of this section discusses the proposed approach to analyzing the data to meet both study objectives, and includes plans for assessing the consequences of omitting some sampled programs—those in saturation markets or that are unable to conduct random assignment—from the experiment.

8.1 Basic Impact Estimates

Beginning with the overall impact estimates, one of the important advantages of a randomized experiment is that it greatly simplifies the data analysis task. For both child- and parent-level outcomes, a simple comparison of means for the treatment and control groups will provide an unbiased estimate of the program's impact. This can be done at the end of each wave of data collection (e.g., at the end of the Head Start year for 4-year-olds). This sequence of comparisons will provide a clear picture of initial program effects and the extent to which impacts are sustained as children enter the early years of primary school.

Although such simple comparisons of group means are a valid analytical approach, the reliability of the impact estimates can, and will, be substantially improved by the use of multivariate regression that includes statistical controls (i.e., independent variables) that are measured **at the time of random assignment**. These control variables will include basic demographic measures related to the individual child (gender, race/ethnicity), the child's family (income, parent's education and employment status, and household composition), and an indicator of which study group the child was randomly assigned to (i.e., 1=treatment, 0=control). The estimated coefficient on this treatment group variable then provides the desired estimate of program impact for the respective dependent variables. In addition, by adding "interaction" terms to these basic impact models we will also be able to examine the extent to which estimated impacts are related to particular baseline characteristics of children and their families (e.g., "Are impacts higher or lower for Hispanic children?").

Once we reach the end of the data collection phase (i.e., at the end of the 1st grade measurement point), we will have sufficient observations to extend the annual impact analyses to include growth curve analysis[49] that can expand the explanatory power of the statistical model by the virtue of having multiple observation points for each child. These models will be estimated for each of the child development measures for which growth over time is an appropriate concept. The actual growth curve analysis will be conducted using "hierarchical linear models (HLM)" that are a relatively new statistical development[50] that properly accounts for the fact that our proposed sampling plan has clustering and its associated correlation between children in the same center, grantee, and community as well as correlation over time (within-child correlation). In estimating these models, an important consideration will be the appropriate form of the growth curve as children do not follow smooth linear trajectories. As a consequence, we will have to explore alternative forms of the growth model to account for expected non-linear growth patterns.

To examine the variation in program impacts as a function of program characteristics, a similar regression model will be estimated that includes additional independent variables accounting for differences in program "models" or other structural characteristics. As with child- and family-level measures, these program measures will be interacted with the indicator of treatment group status to assess the extent to which impacts are different for children who attend different types of Head Start programs. In addition, once all waves of data collection are complete, similar HLM growth models will be estimated that include "levels" that use program characteristics to assess the extent to which differences in Head Start operations affect children's growth and development.

Similar models can be used for each outcome of interest, although the independent variables included in the regression or HLM model may vary from outcome to outcome. We will also run tests of the statistical significance of each measured effect, tests whose ability to detect impacts when impacts occur (i.e., whose "power') was discussed in an earlier section of this document. Tests for differences in impacts across subgroups-subgroups defined by child, family, program, and/or contextual characteristics—will also be run. However, we will **not be able to test for effects in individual sites** (i.e., individual grantees, delegate agencies, or centers) given our relatively small site-specific data samples.

8.2. Impacts On The "Treated"

Not all children assigned to the treatment group will receive Head Start or "Head Start-like" services, while inevitably some of those assigned to the control group will. As a result, the basic treatment-control comparisons just described will not measure completely the average impact of the Head Start treatment. Rather, they will show the impact of being **assigned** to the treatment group. This is an important and useful measure for reasons discussed earlier in conjunction with our treatment definition. It describes how the outcomes of children granted access to federally-funded Head Start services differ from those of comparable children eligible for similarly comprehensive programs with independent funding where they exist.

---

[49] Singer, J. (1999). "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." ***Journal of Educational and Behavioral Statistics***, 24(4), pp. 323-355.

[50] Bryk, T. & S. Raudenbusch (1992). ***Hierarchical Linear Models: Applications and Data Analysis Methods***. Newbury Park, CA: Sage Publications, Inc.

There is, however, also an interest in the effect of Head Start on just "the treated," i.e., those children who actually receive the comprehensive service package defined by the Head Start performance standards. Indeed, it is the comparison of this group to an otherwise similar set of children not receiving comprehensive services that provides data on the preferred impact question identified above: What difference do services meeting the Head Start performance standards make for child outcomes relative to a (hypothetical) world where such comprehensive services do not exist? In sites where all comprehensive service programs are funded at least in part by Head Start, we can create this alternative world by assigning selected Head Start applicants to the control group, where they are precluded from accessing services funded by federal Head Start-dollars. But not where other state Head Start or "Head Start-like" services are available from exclusively non-Head Start sources (e.g., comprehensive state pre-K programs), a circumstance in which the alternative provider may not even participate in the Head Start program. As discussed previously, there is no feasible way to keep control group members from getting these services, possibly from the very agency that assigned them to the control group but in a non-Head Start center or classroom.

Even in sites where all the programs that meet Head Start performance standards receive federal Head Start funding, it is likely that some control group members will participate in fully comprehensive services either by "slipping through the cracks" at the Head Start agency that originally assigned them or, more likely, by turning to other state Head Start providers in the area. Some number of "cross-overs" of this sort is inevitable in a randomized study, despite our attempts to do everything possible to prevent it (see our earlier discussion of monitoring and maintaining random assignment)—possibly going so far as to pursue agreement with other non-evaluation Head Start grantees and delegate agencies to comply with the design and exclude controls where service areas overlap.

Whether by accessing Head Start itself or receiving equivalent services from non-Head Start programs, we will wind up analyzing some control group members who essentially got the "treatment" along with some treatment cases that did not. In a key paper that extends and interprets already well-known methods of dealing with these problems, Angrist et al. (1996)[51] describe a method for deriving accurate measures of the effect of treatment on the "treated." Essentially, their approach--a version of the econometric technique of "instrumental variables" — re-scales the overall difference in observed outcomes between the treatment and control groups into an average effect on just those children who receive different types of services because of randomization. Children assigned to the treatment group who never participate in Head Start — a group called "no-shows" in the evaluation literature[52]— presumably wind up in childcare arrangements similar to their counterparts in the control group; both sets of families are free to choose any other options available in the community. Similarly, we can presume that control group members admitted to Head Start or other sources of equally comprehensive services — a group called "cross-overs" in the literature — have experiences similar to their counterparts in the treatment group. It is only the residual group--the portion of the random assignment population that (a) if assigned to the treatment group, will participate but (b) if assigned to the control group, will not participate — that experiences any differential effects. This share can be estimated as:

---

[51] Angrist, J., G.W. Imbens, & D.B. Rubin (1996) "Identification of Causal Effects Using Instrumental Variables," ***Journal of the American Statistical Association,*** June, 91(434), pp. 444-472.

$$1 - n - c,$$

where "n" is the no-show rate (no-shows/total treatment group members) and "c" is the cross-over rate (cross-over controls/total controls). When the basic impact estimates are divided by this factor, they "expand" into estimates of the effect of treatment on just the treated.[53] In a regression framework, the equivalent result is obtained by specifying the outcome, or dependent, variable as a function of participation in Head Start or Head Start-like services (as opposed to assignment to the treatment group as in the usual regression approach) and then using assignment to the treatment group as an "instrumental variable" when estimating the coefficient on the participation variable.

If the assumptions of no effect on "no-shows" and of identical impacts on "cross-overs" and "cross-over-like" members of the treatment group are correct, this adjusted estimate of the effect of treatment on the treated is just as reliable as the original impact estimate for all assigned. Indeed, the same test for statistical significance applies to both estimates. This technique is not a panacea, however, since the absence of comprehensive services for a "no-show" child does not necessarily imply exact correspondence with a control group case. For example, the mere expectation of participation in Head Start may lead some "no-show" families to pass up other care arrangements for their children in the interim between random assignment and the scheduled start of services that equivalent families in the control group pursue, opening a "wedge" between the experiences and--potentially--the outcomes of the two groups. Similarly, a "cross-over" and her/his treatment group counterpart do not necessarily receive the exact same services in a comprehensive Head Start or Head Start-like program when one accesses services "through the front door" and the other has to take a more indirect and possibly longer-run route to the "treatment" by skirting random assignment or going to an alternative Head Start agency.

---

[52] See Bloom, H.S. (1984), "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 1984, Vol. 8 (April), pp. 225-246.

[53] Formally, any initial impact estimate represents the average impact of assignment to the treatment group across all treatment group children.. This can be expressed as a weighted average of three separate impacts, one for children in the "no-show" population, one for children in the "cross-over" population, and one for all other children. Given their initial equivalence through random assignment and their parallel program experiences, treatment and control group children in the first of these populations--"no-shows" and "no-show-like" controls—can be expected to have similar outcomes. Thus, on average they will experience an impact of 0. The same argument can be made regarding treatment and control group children in the second of these populations—the "cross-overs" and "cross-over-like" members of the treatment group. Hence, the overall impact estimate, E, can be expressed as the weighted average of two 0 effects plus a potentially positive effect on the remaining subpopulation of "non-no-show-prone, non-cross-over-prone" children. Using F to represent this latter quantity, $E = (n)0 + (c)0 + (1-n-c)F$. The first two terms drop out, leaving $E = (1-n-c)F$, or $F = E / (1-n-c)$.

In providing estimates of effects for both all treatment group members and just the "treated," clear labeling and interpretation of the results becomes paramount. In sites where comprehensive "Head Start-like" services reach an important share of the control group, we must make clear that the basic impact estimates—prior to the "no-show" and "cross-over" adjustments--do not reflect the full value of the Head Start service package to all who receive it.

The same points carry over to national estimates that incorporate data from sites of this sort: overall, the basic, unadjusted national estimates will **understate** the average value of what federal Head Start pays for across the country by in some instances comparing it to essentially equivalent services. Separate estimates may be needed to represent the part of the country where few if any "Head Start-like" services are available that are not federally funded and, hence, the basic unadjusted estimates are fine, and the part of the country where such services do exist in important quantities and these caveats apply. It is in anticipation of this need that we have made the extent of alternative State Head Start or "look-alike" services in the community a major stratifier in choosing the research sample. It is also the reason we attach such importance to adjusted impact estimates that—on an only partially experimental basis—measure the impact of Head Start services relative to other, less-intensive, service options in **all** sites, and therefore for the Nation as a whole. In choosing between the two ways of looking at the effect of treatment on the "treated," our interpretation of results will need to recognize the trade-off posed between purely experimental estimates in the "non-look alike" sites, which have strong internal validity but may not represent the nation well overall and adjusted estimates for a nationally representative sample that assure external validity at some risk to internal validity. The ability of the design to address the question from both perspectives should strengthen our hand in making sure to draw the right conclusions on this critical impact question.

The analysis of the second major question underlying the study—Under what circumstances, and for which children and families, does Head Start work best?—is less sensitive to these issues. We can rely exclusively on the basic, unadjusted estimates for that purpose and know we have complete internal validity. It may make sense to confine this analysis to just those sites with few if any "look-alike" programs, depending on how many Head Start centers this would remove from the "pool" of local program approaches across which variations in impact can be examined. The alternative is to make sure to include the availability of "look-alike" services as one of the key contextual variables used to explain impact variations across centers. In making this choice, we need not worry much about external validity, since remaining nationally representative is not essential here with the focus on **variations** in impact rather than the population wide average impact.

8.3. Checking for "Non-Response" Bias in the Experimental Estimates Due to Grantee Exclusion

As noted in the discussion of our sampling plan, two types of grantees and delegate agencies will be excluded from the random assignment experiment:

- Grantees and delegate agencies operating substantially below capacity—or at capacity in saturated market—where there are not enough eligible families seeking Head Start services to place some in a control group without reducing the total number served; and

- Grantees and delegate agencies that have not saturated their "markets" and could identify enough cases to form a control group, but who cannot be convinced to implement random assignment.

Our plan for dealing with this component of the national Head Start program in the analysis is quite similar for the two subsets, though there are some important differences with regard to data collection. Because these "missing programs" could affect the representativeness—or external validity—of all of our impact estimates, we need to determine the extent to which they differ systematically from included programs. The risk of analyzing a skewed sample in the experiment is exactly analogous to that posed by non-response bias on a household survey, where one must consider the possibility that missing respondents represent an important--and distinctive--part of the overall picture one wants to convey. This analogy provides a useful framework for thinking about what may be missing from the experimental results and how to accommodate this limitation in the analysis.

A number of strategies are available for examining exclusions from the experiment, just as there are many ways to analyze non-response bias in survey data collection. The most promising strategies evident at this point in the design process are listed in Exhibit 8. All but the first represent extensions beyond standard practice in examining the reliability of randomized experiments for generating nationally representative findings. The first three—performing "background" checks, simulating potential bias, and examining programs that "just barely" made the experimental sample—parallel techniques used to check for non-response bias in survey data collection, illustrating the power of the "non-response bias" framework in attacking the problem of "missing" experimental estimates.

While innovative and extremely valuable, these initial strategies are not in our view sufficient to guard the Head Start impact study fully from criticism that it is not nationally representative—i.e., that it lacks external validity. Other strong, multi-site experiments have been tellingly challenged on this basis, despite their many other strengths. It would be extremely unfortunate if ACYF's Head Start evaluation fell prey to this suspicion—particularly when other, even more powerful techniques for dealing with omitted experimental sites are available from the literature. We propose to use these tools—particularly the last two items listed in the exhibit—to strengthen the case for the national validity of the Head Start experiment. Each is described below.

Exhibit 8: Analysis Options for Examining Non-Random Assignment Programs

- **Performing "Background" Checks**
  Compare centers and grantees on background factors to see if those without random assignment differ systematically from those with random assignment in terms of their service approaches, participant characteristics, and/or non-Head Start services available in the community. Closer correspondence on these factors, which collectively determine Head Start impacts, reduces the threat of bias in the main impact results.

- **Simulating Potential Bias**
  Do a sensitivity analysis of the main impact findings from the random assignment programs to determine their possible bias by assuming different levels of impact for the non-random assignment programs in the sample and then recalculating national findings under each scenario. If scenarios differ little (e.g., as when there are not very many non-random assignment cases) conclude that omission of non-random assignment programs from the main results do no appreciably bias the findings.

- **Examining Programs that "Just Barely" Made the Experimental Sample**
  Concentrating on the experimental sample, compare estimated impacts between random assignment programs that easily met the requirements for inclusion in the experiment (operating at capacity, willing to conduct random assignment) and those that met the conditions but only marginally. If the two groups do not differ appreciably in their Head Start impacts, more "extreme" cases where random assignment was entirely impossible are less likely to do so.

- **Validating Non-Experimental Approaches Using the Experimental Results**
  Estimate impacts for the non-random assignment programs using standard non-experimental methodologies. Then test the validity of each method by checking if it gives the right answer for the programs that did do random assignment, using the experimental impact estimate as the "gold standard."

- **Using Non-Standard Non-Experimental Approaches**
  Draw from recent literature on innovate non-experimental methods to strengthen comparisons with the experimental findings, including the "regression discontinuity" approach, "internal" comparison sites, and sibling models [see text].

Validating Non-Experimental Approaches using Experimental Results. A considerable literature has sprung up over the last 15 years on validating non-experimental methods for calculating program impacts using experimental findings as the benchmark. The idea is simple but powerful: when wondering whether to trust a non-experimental approach, find a place where randomized experiments have been run and see if it works as well as the experimental approach. This strategy has been pioneered by LaLonde (1986), Friedlander and Robins (1992), and Bell, et al. (1995) in the area of adult employment and training programs[54] but can be applied equally to other social program areas, including Head Start.

Here, the application takes a new twist. Rather than seeing non-experimental methods as an alternative to random assignment in future evaluations—the usual perspective in the literature—we will focus on testing the validity of those methods as a supplement to random assignment. If a method or methods can be found for reliably calculating Head Start's impact non-experimentally in the non-random assignment sites, it can serve as a check on the reliability of the main experimental results as a representation of the nation. The non-experimental estimation approaches that might be considered in this role include:

- examination of changes in Head Start participants' school readiness and social competencies over time as indicators of program impact (the "pre/post" approach);

- pre/post comparisons that control for national norms in child development over the same period, such as average gains in PPVT test scores;

- comparison group options that use data on non-Head Start children from the same communities as the non-random assignment Head Start programs to approximate what would have happened to participants absent the program, calculating program impact as the difference between participant and comparison group outcomes;

- statistical matching of Head Start participants with non-participants on background characteristics in forming comparison groups; and

- econometric controls for observable and, through "selection modeling" techniques, unobservable difference between participants and comparison group members.

We have budgeted for collection of data in the non-random assignment sites to support the first two of these analyses, both baseline (pre) and follow-up (post) data for a substantial number of children participating in Head Start in non-random assignment programs and their families. FACES 2000 data (see Appendix B) will supply information on grantees and delegate agencies in saturation communities that are ineligible for inclusion in the experiment. New primary data collection will be conducted in sites dropped from the experiment because a grantee or delegate agency refused to agree to random assignment.

---

[54] Robert J. LaLonde, "Evaluating the Evaluations of Training Programs with Experimental Data," American Economic Review, 1986, provided the first example of this approach. Daniel Friedlander and Philip K. Robins, "Estimating the Effects of Employment and Training Programs: An Assessment of Nonexperimental Techniques," unpublished paper from the Manpower Demonstration Research Corporation, 1992, extended it to comparison site models for non-experimental impact estimation. Stephen H. Bell et. al., Program Applicants as a Comparison Group in Evaluating Training Programs, W.E. Upjohn Institute for Employment Research, 1995, first used it to look at "internal" comparison groups of unserved program applicants.

We will use each of the available non-experimental methods to calculate impacts separately for both the non-random assignment and random assignment programs in the study. The latter—when compared to the experimental results—will provide a check on the reliability of the former: if a technique works well in replicating the experimental finding for the random assignment sample, we would expect it to succeed in estimating effects in other sites where the experimental benchmark is not available.[55] We will rely on any non-experimental method that meets this test for measures of Head Start's impact in circumstances not covered by the experiment—both to examine the degree of "omitted program bias" possible in the main, purely experimental study results and, if appropriate, to provide adjusted estimates that correct this bias.

The "Regression Discontinuity" Approach, "Internal" Comparison Sites, and Sibling Models. The same validation procedures can be applied to more innovative non-experimental estimation methods than just those conventionally used in the literature. As many as three such procedures will be considered for the Head Start evaluation depending on data availability: "regression discontinuity" analysis, comparison site designs, and sibling models.

The ranking of Head Start applicants with "scores" reflecting local targeting criteria for the program—and the use of these rankings for admission purposes—provides a perfect opportunity for applying the **"regression discontinuity"** approach to estimating program effects for the non-random assignment sample. Developed in the education field and introduced to the econometric literature on training program evaluation by Bell et al. (1995, op cit.), this approach looks for a discrete jump in outcome levels at the cut-off point for admission on the ranking scale. If Head Start participants just above the cut-off achieve better outcomes than unsuccessful applicants just below the cut-point, the "discontinuous" jump in outcomes can be interpreted as the program's impact on the most marginal participants. The method can be extended to effects on all participants using regression analysis which expresses outcomes as a linear function of the targeting "score" and a dummy variable for program participation; the extension of the "without-program" regression line to scores above the admission point provides a benchmark for calculating impacts for participants of all types. Subject to data availability—something to be explored during the process of recruiting sites—this method could be of particular value for programs that have more applicants than they can serve (i.e., those in non-saturation sites) but do not agree to conduct random assignment.

---

[55] A crucial consideration in applying this technique is the determination of when a non-experimental method "works well" compared to the "right" answer provided by the experiment. Though often overlooked in the literature, this is a complex statistical matter sorted out by Stephen H. Bell and Larry L. Orr in "Are Non-Experimental Estimates Close Enough for Policy Purposes? A Test for Selection Bias Using Experimental Data," Proceedings of the Section on Government Statistics of the American Statistical Association, 1995. Bell and Orr provide an explicit standard for making this determination--i.e., for deciding which non-experimental methods to trust--and illustrate its application to a set of job training experiments.

Another non-experimental means of gauging the impact of non-random assignment programs uses control group members from the random assignment sites to represent the "counterfactual" for Head Start participants in the non-experimental sites. Except for possible differences in community environment, these controls should closely match Head Start participants from other locations after controlling for measured differences in family background and pre-program development. The match should be especially strong on the often difficult-to-measure factors that influence the decision to apply to Head Start, since both sets of families will have taken this step. This strategy—first devised and tested by Friedlander and Robins (1992, op cit.) in job training evaluation—treats control group members as **a comparison site sample** of similar children and families in places where the program does not exist, but one with the especially attractive feature of having equal interest in Head Start services as participants. A variety of such strategies (each relying on a different subset of random assignment programs for its control group data) can be tried under this approach. Those found reliable compared to experimental results can then be used in non-experimental sites without incurring the cost of any additional data collection or analysis file development. We view this innovation as a very efficient way to explore additional non-experimental analysis approaches for the Head Start evaluation.

Another promising way of developing and testing impact estimates for the non-random assignment programs in the evaluation comes from recent analyses by Janet Currie and Duncan Thomas,[56] who estimated the national impact of Head Start from existing survey data on Head Start children and their siblings. These innovative **"sibling" models** capitalize on a sort of natural experiment in which one child in a family participates in Head Start while another of similar age does not. Using these non-Head Start siblings as a natural "control group," Currie and Thomas were able to produce highly convincing measures of Head Start's impacts without the benefit of an experiment. If similar data and additional analysis resources become available to the current national evaluation, we can apply the method to the non-random assignment programs sampled for the evaluation and, for the first time, test it against the experimental norm by applying it to the experimental portion of the sample.

8.4. Measuring Community-Wide Effects

Head Start is hypothesized to have at least one community-wide benefit that cannot be captured within the context of the proposed study: the influence of Head Start on "raising the bar" for preschool day care services generally within the community. If the presence of Head Start services, and particularly its emphasis on school readiness activities, changed the expectations of parents and/or the objectives of caregivers in non-Head Start programs, child development efforts and results in the preschool years may have risen community-wide making Head Start responsible for benefits to children and families beyond those it serves directly. In the experiment, these benefits could actually count against the program by raising outcomes for control group children participating in other preschool programs.

---

[56] See, for example, Janet Currie and Duncan Thomas, "Does Head Start Make a Difference?", American Economic Review, June 1995.

From the standpoint of the evaluation, these benefits can be viewed in one of two ways, guided by the concept of "market leadership" put forth by economists:

- As an historical contribution of Head Start that no longer depends on the continuation of Head Start as such. Under this view, community norms and provider practices will continue at the heightened level induced by the entry of Head Start into the "marketplace" for child care many years ago, even if the program itself some day "goes out of business." An evaluation undertaken to guide *future* decisions on Head Start policy and funding need not take these benefits into account, since they will continue regardless of the policy decisions made now or in the future. If these heightened norms make it difficult for Head Start today to achieve better results than other community alternatives—the "counterfactual" to be represented by the control group—its past success indeed reduces its current and future value, a factor appropriately reflected in the experimental impact estimates.

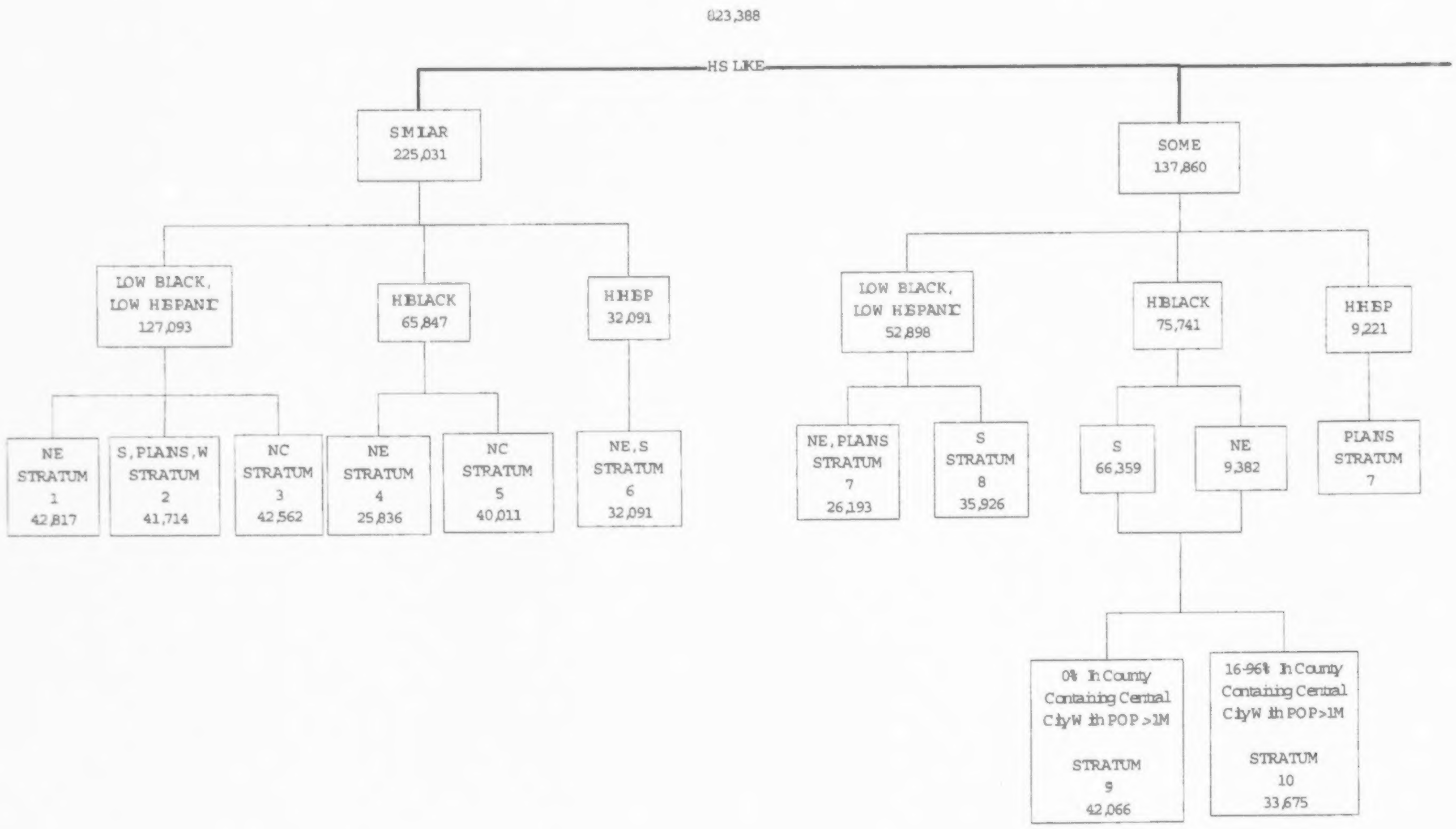
- Alternatively, it may be that Head Start's *continued* presence in the "market" is important in maintaining higher community norms and the value of child care generally. This seems especially likely in communities where school-readiness techniques continue to evolve and improve due to Head Start's lead, although the departure of Head Start from an otherwise stable marketplace could lead to a decline in pre-school child care practice and value generally.

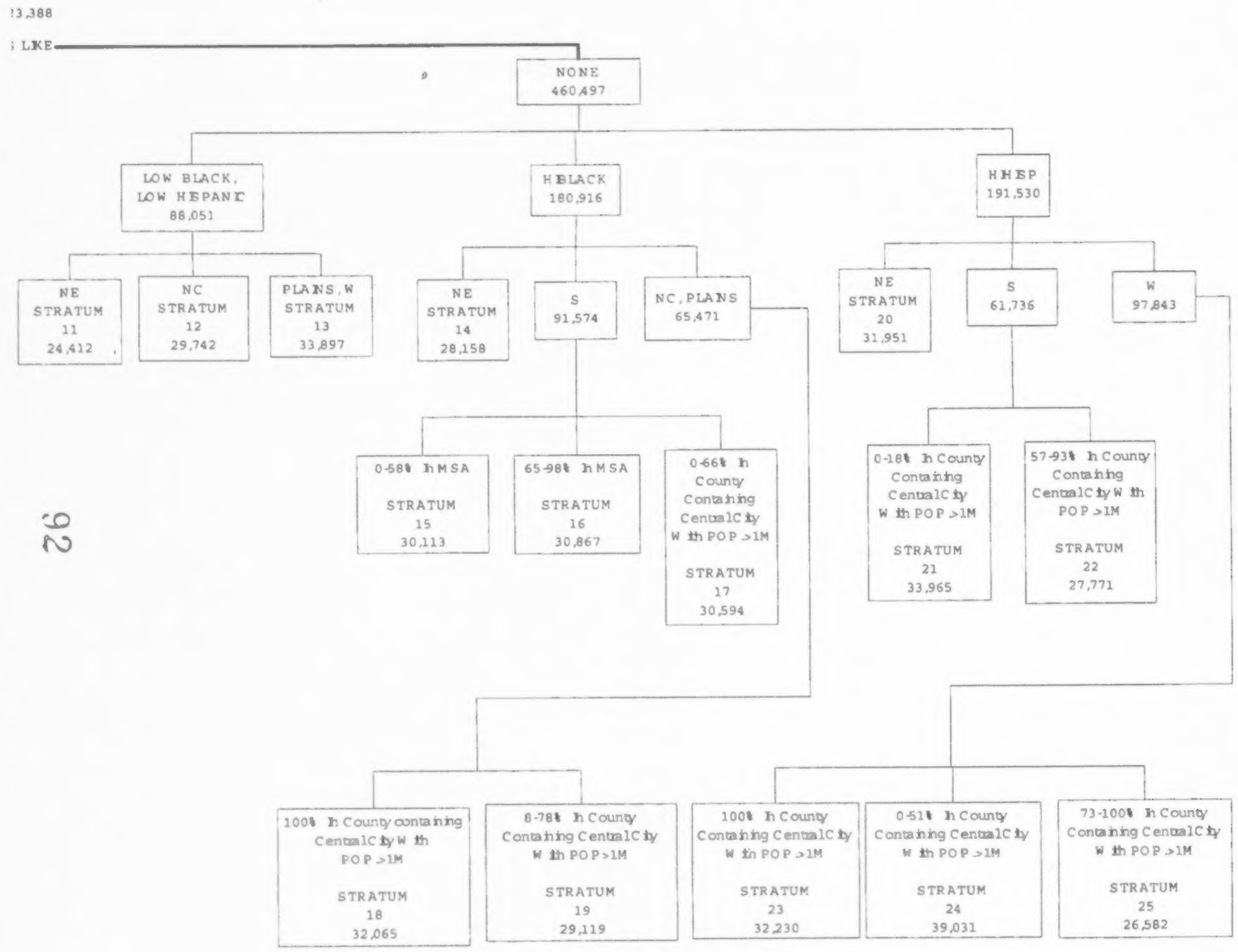
Our qualitative analyses and interviews will explore this latter possibility with community leaders both in and outside Head Start. If Head Start's "market leadership" continues to matter to child care practice more broadly in the eyes of these experts, this fact needs to be reported as a possible unmeasured benefit of the program into the future.

Appendix A: Sample Cluster Stratification

See attached 2-page diagram.

NATIONAL HEADSTART IMPACT STUDY STRATA FOR CLUSTERS


823,388
HS LIKE
SIMILAR 225,031
SOME 137,860
LOW BLACK, LOW HISPANIC 127,093
HI BLACK 65,847
HI HISP 32,091
LOW BLACK, LOW HISPANIC 52,898
HI BLACK 75,741
HI HISP 9,221
NE STRATUM 1 42,817
S, PLAINS, W STRATUM 2 41,714
NC STRATUM 3 42,562
NE STRATUM 4 25,836
NC STRATUM 5 40,011
NE, S STRATUM 6 32,091
NE, PLAINS STRATUM 7 26,193
S STRATUM 8 35,926
S 66,359
NE 9,382
PLAINS STRATUM 7
0% In County Containing Central City With POP >1M
STRATUM 9 42,066
16-96% In County Containing Central City With POP>1M
STRATUM 10 33,675


3,388
LIKE
NONE 460,497
LOW BLACK, LOW HISPANIC 88,051
HI BLACK 180,916
HI HISP 191,530
NE STRATUM 11 24,412
NC STRATUM 12 29,742
PLAINS, W STRATUM 13 33,897
NE STRATUM 14 28,158
S 91,574
NC, PLAINS 65,471
NE STRATUM 20 31,951
S 61,736
W 97,843
0-58% In MSA STRATUM 15 30,113
65-98% In MSA STRATUM 16 30,867
0-66% In County Containing Central City With POP >1M STRATUM 17 30,594
0-18% In County Containing Central City With POP >1M STRATUM 21 33,965
57-93% In County Containing Central City With POP >1M STRATUM 22 27,771
100% In County containing Central City With POP >1M STRATUM 18 32,065
8-78% In County Containing Central City With POP >1M STRATUM 19 29,119
100% In County Containing Central City With POP >1M STRATUM 23 32,230
0-51% In County Containing Central City With POP >1M STRATUM 24 39,031
73-100% In County Containing Central City With POP >1M STRATUM 25 26,582

Appendix B: Ways to Analyze Impacts without Placing a Two-Year Exclusion on Controls

Suppose randomization to the control group meant just a single year of exclusion from Head Start, not a full two-year embargo? Families of children accepted for admission to a two-year sequence (i.e., newly-entering 3-year olds) would still be served even if they were assigned to the control group at program entry, but services would be deferred for a year. In contrast, those assigned to the treatment group would be enrolled immediately and allowed to participate for the full two years.

This design option provides a direct measure of the impact of Head Start participation at age 3 years as an addition to the "standard" 1-year of services starting at age 4 years:

(1) Impact of E given S = (average outcome for Ts) - (average outcome for Ds) ,

where E represents the early, age 3, year of Head Start services and S the standard year, and where Ts and Ds are members of the treatment group and the "deferred access" control group respectively. Since both Ts and Ds participate in Head Start at age 4 and—through random assignment—are similar in background characteristics, the difference in average outcomes between the two groups reflects the impact of the earlier year of service (received by the treatment group alone) as a supplement to the standard year (received by all sample members). The same basic model can be used to obtain a measure of the impact of the "standard" year of Head Start participation as a "stand alone" option—and of a combined two years of participation. But here the derivation, and the results, are not so straightforward.

Notation and Potential Experimental Groups

Some additional notation will aid our analysis of the extra, more complex cases. Consider two additional groups of children existing for the moment only hypothetically:

Cs = children denied access to Head Start for the full two years (true controls)

Rs = children served by Head Start initially (i.e., during the early year that begins at age 3) but then removed from the program prior to the start of the pre-K year.

A matrix (shown on the next page) will help us keep track of the four experimental groups introduced to date, defined by the extent and timing of their Head Start program exposure.

Group	Participate in Head Start during Early Year (E)	Participate in Head Start during Standard Year (S)	Observed in the Data?
T [treatment]	X	X	yes
R ["removal"]	X	-	maybe
D ["deferral"]	-	X	yes
C [control]	-	-	no

It will also be useful to simplify notation for impacts and outcomes, restating Equation 1 as

(1') $\text{Impact } E \mid S = Y(T) - Y(D)$,

where Y(□) represents the average outcome of the group in parentheses. Having measured the impact of the early year *as an addition to the standard year* in this way, we are also interested in the impact of a full two years of Head Start participation, the early plus standard years:

(2) $\text{Impact } E + S = Y(T) - Y(C)$.

Finally, the impact of just the standard year is of some interest, both for the population of children and families now provided with two years of service and those only seeking (or only obtaining) one year.

The Challenge and a First Response

We want to obtain all these results without creating or observing Y(C), outcomes for a fully excluded control group. This is not a problem in the first year of the analysis, when the deferral group, D, looks just like a pure control group (see matrix). Neither group participates in Head Start during that year, so Y(T) - Y(C) can be calculated without bias by substituting Y(D) for Y(C) and calculating Y(T) - Y(D). However, this expedient is not available in the second and subsequent years of the research period once the deferral group D becomes "contaminated" by participating in the standard year of Head Start services. By this point, all of the three potentially observable groups—T, R (which we may or may not observe), and D—have spent at least some time in the program.

One way around this dilemma is to assume "strict additivity" of impacts across the two years of program participation. In any one year, starting with the year of "standard" Head Start participation and continuing to the end of the follow-up period (the end of 1st grade), the treatment group T may experience improved outcomes relative to a pure control group C for two reasons:

- Ts may benefit from the Head Start assistance received during the earlier of the two years of participation as benefits of the initial "step up" to faster development continue to accrue throughout a child's early elementary years.

- Head Start assistance received during the standard year of participation will also have "carry-over" benefits to succeeding years.

While both gains—those engendered by the early Head Start year and those engendered by the standard year—could occur at the same time for the same child, conceptually they can be kept separate. The question is whether the two gains interact with one another when both are present, making Head Start's total impact greater than (or, with negative synergism, less than) the sum of the parts.

To explore this option, we will start with the much simpler case of only a single year of Head Start enrollment. We can think about the impact of an early Head Start year (among those who get it) alone—i.e., *absent* enrollment during the standard year. We can represent this case as follows:

$$(3) \quad \text{Impact } E \mid \text{no } S = Y(R) - Y(C),$$

since Rs participate in Head Start only in the earlier year and Cs not at all.

Similarly, we can think of the impact of the standard year of Head Start *alone* over that same period—i.e., absent participation in the *early* Head Start year:

$$(4) \quad \text{Impact } S \mid \text{no } E = Y(D) - Y(C).$$

Strict additivity says that the impact of these two years of Head Start participation, when combined as a "package," equals the sum of the two individual effects, no more and no less:

$$(5) \quad Y(T) - Y(C) = [Y(R) - Y(C)] + [Y(D) - Y(C)].$$

This formulation assumes a complete lack of synergism between the two years of Head Start services: the benefit a child receives from an early year of Head Start enrollment is the same regardless of whether a second year is added, and the benefit from the standard year is the same whether or not it was preceded by an early year. No third term showing interaction or synergism between these components, either positive or negative, appears on the right-hand side of equation 5, leaving the whole exactly equal to the sum of the parts.

Equation 5 converts into

$$(5') \quad Y(T) = Y(R) + Y(D) - Y(C)\,,$$ which implies in turn that

$$(6) \quad Y(C) = Y(R) + Y(D) - Y(T)\,.$$

Substituting back into equation 2, we get

$$(2') \quad \text{Impact } E + S = Y(T) - [Y(R) + Y(D) - Y(T)]$$

$$= 2\,Y(T) - Y(R) - Y(D)\,,$$

which can be calculated without using a pure control group as long as the other three types of "exposure" to the Head Start treatment are created—full treatment for Ts, an initial year only for the "removal" group R, and the later standard year only for the "deferral" group D.

Some Caveats

Despite its advantage in eliminating the need for a pure control group, there are several reasons why we might hesitate to use equation (2') to measure Head Start's impact:

- Execution of this approach requires 3-way random assignment of children applying for the early year of Head Start into separate T, R, and D groups. This complicates both the administration of random assignment and subsequent program actions with respect to children assigned to different study groups. Perhaps more importantly, it substantially increases the sampling variability of all impact estimates, both because (for a given total sample size) each study group grows smaller by one third to accommodate three rather than two groups, and because the impact formula now contains three independent terms—one even multiplied by a factor of 2—effectively quadrupling the variance of the impact estimate and thus doubling the size of the effects that can be detected with confidence.

- It also means pulling some children—the Rs—out of Head Start after just one year of what is intended to be a two-year sequence. This could have two detrimental effects, adding a new potential concern regarding the ethics of random assignment (is it better to start services and then interrupt them, or to not start them at all?) and potentially engendering distorted behavior on the part of the families and/or Head Start workers who deal with the "one year and out" children. For example, knowing a child cannot stay in Head Start a full two years, a family may put the child in a non-Head Start care arrangement right away, effectively turning what was to have been an R case into a C case who receives no Head Start services at all. Similarly, Head Start staff might invest less in a young child known to be leaving the program at the end of the year than would otherwise be the case—or invest more to push his/her school readiness ahead faster in the limited time available.[57]

- The assumption of strict additivity of effects may be wrong, depending on what one believes about complementarities between successive years of Head Start exposure. For example, one might think that the benefits of receiving the standard year of Head Start at age 4 go up if the child previously received Head Start services at age 3 to "lay the foundation" for greater gains in subsequent years. Alternatively, Head Start services may—like many other products and services in the economy—yield diminishing returns as one adds layer upon layer, making the standard year of service most productive when it is the *only* year of services.

This last concern not only points to the possibility that impact estimates based on equation (2') are biased, it also illustrates the difficulty one has under this approach judging the likely *direction* of bias. If an earlier year's enrollment makes the standard year's assistance more effective, equation (5) should have a third, positive term on the right-hand side and equation (2') underestimates the overall contribution of two years of Head Start enrollment by leaving out the positive synergism. If instead the standard year of Head Start becomes less valuable in its net contribution once a child has the initial, early year of Head Start services under her/his belt, equation (5) is missing a negative term on the right-hand side and equation (2') overestimates the contribution of the two years together by allocating full value to each year, including twice the benefit of an "initial dose" that in practice will only occur once.

A Lower-Bound Strategy

An alternative approach to estimating overall effects without a pure control group focuses on bounding the likely impact above and below as a way of narrowing our uncertainty regarding the direction and degree of bias. Two additional estimates of Head Start's impact come into play at this point, one likely biased upward and the other likely biased downward. While neither of the estimates is likely to provide the "right" answer, jointly they provide useful upper and lower bounds on that quantity. Hence, their advantage is not so much that we think them unbiased but that the direction of bias is known or can be assumed with some confidence.

---

[57] For legal and ethical reasons, it is probably not an option to delay the identification of Rs until the end of the first Head Start year, grouping Rs and Ts together (as simply "program participants") during the first year. The informed consent form signed by applicant families prior to any randomization needs to indicate all the randomization that will take place. This advance announcement of a later, second round of assignment—and the "double jeopardy" it suggests—may have just as chilling an effect on family participation in Head Start as would telling a subset that the initial year will definitely be the only one.

Beginning in the second year of Head Start enrollment and for all subsequent years, a lower bound on program impact can be calculated as:

$$\text{(7)}\quad \text{Lower bound (Impact E + S)} \;=\; [\, Y(T) - Y(D) \,] \;+\; [\, Y(T^*) - Y(C^*) \,]\;,$$

where T^* and C^* are the treatment and control groups created *from children who apply for Head Start for the first time in the later, standard year, just a year before kindergarten entry*. This latter set of children has not been considered to this point but will be an essential part of the overall evaluation, providing measures of the impact of the one-year version of Head Start for children whose families seek only a single year (or who are only admitted at the later point having applied but been excluded the previous year) The question of two years of artificial exclusion from the program never comes up for this group, so it is not directly germane to the topic of this section. It can be analyzed experimentally using standard tools each year [i.e., impact $S = Y(T^*) - Y(C^*)$]. We simply "import" that calculation here to help us deal with the thorny problems of the two-year program—specifically by approximating a parallel impact measure for the two-year service population.

To see how this is done, return again to equation (2) but now expand it by adding and subtracting a common term::

$$\text{(2'')}\quad \text{Impact E + S} \;=\; [\, Y(T) - Y(D) \,] \;+\; [\, Y(D) - Y(C) \,]\,.$$

Here, $Y(D) - Y(C)$ represents the impact of Head Start services that start just a year prior to kindergarten *for children who normally enter the program a year earlier*. Not observing $Y(C)$, we don't know what this quantity is. What we do know is the corresponding quantity for a somewhat different population, *children who normally enter Head Start only a year before kindergarten,* $Y(T^*) - Y(C^*)$. The two expressions match up because the children in T^* follow the same sequence of events as the children in D—no Head Start participation at age 3, followed by Head Start participation at age 4—and the children in C^* mirror those in C (no Head Start participation at any time). Importantly, however, the two groups of children are not the same in their underlying characteristics, nor do they necessarily come from the same kinds of homes nor receive the same exact treatment from Head Start.

The assumption required to use $Y(T^*) - Y(C^*)$ in place of $Y(D) - Y(C)$ in forming a lower bound is that

> *The children who can benefit most from Head Start are the most likely to enter the program early (i.e., at around age 3 rather than around age 4).*

This assumption will hold as long as two other seemingly plausible conditions are met:

- Parents, or Head Start intake staff, or both put greater emphasis on early Head Start enrollment for children they believe will benefit most from participation.

- The same parents and/or staff have at least some ability to judge reliably which children in fact *will* benefit most from Head Start participation.

When these conditions hold early entrants experience greater initial impacts than older entrants and

$$(8)\quad Y(D) - Y(C) \; > \; Y(T^*) - Y(C^*) .$$

Substituting the smaller of these expressions for the larger in equation (2''), we get

$$\begin{aligned}(2''')\quad \text{Impact E + S} &= [\,Y(T) - Y(D)\,] + [\,Y(D) - Y(C)\,] \\ &> [\,Y(T) - Y(D)\,] + [\,Y(T^*) - Y(C^*)\,] ,\end{aligned}$$

showing the estimate in equation (7)—the right-hand side of this new equation—to be a lower bound on the true impact of two years of Head Start enrollment for the children normally served for that long. Equally important, all of the terms in equation (7) can be calculated from an experiment based on simple 2-way random assignment to a full treatment group (T) or a deferred treatment group (D)—for children who normally enter Head Start two years before kindergarten—and a treatment group (T^*) or control group (C^*)—for children who normally enter one year before kindergarten.

A Complementary Upper Bound

To complement this lower bound, we can *from the same experiment* calculate an upper bound on Head Start's impact on two-year participants during the second year of participation:

$$(9)\quad \text{Upper bound (Impact E + S)} = [Y(T) - Y(D)\,] + [\,Y^{\wedge}(T) - Y^{\wedge}(D)\,] ,$$

where $Y^{\wedge}(\cdot)$ indicates the average outcome for the group in parentheses a year prior to the standard, second year of Head Start participation and $Y(\cdot)$ again represents average outcomes during the standard, second year. Interestingly, equation (9) sums the difference in outcomes between children allowed to enter Head Start at age 3 (the Ts) and those deferred until age 4 (the Ds) across two successive years of observation. For this "sum-of-differences" estimate to be upper bound on Impact E + S, we must assume that

> *A given set of Head Start services has a larger immediate (i.e., same-year) effect on a given child if begun earlier in that child's life.*

This assumption accords with recent research on the relative importance of developmental inputs at different stages of a child's life, and when applied to the two years leading up to kindergarten entry implies that:

$$(10)\quad Y^\wedge(T) - Y^\wedge(C) > Y(D) - Y(C),$$

Here, the left-hand term shows the benefits of Head Start if participation begins two years prior to kindergarten entry and the right-hand side shows the benefits—for the same population—of Head Start participation begun a year later. Substituting the larger of these two for the smaller in equation (2"), we get

$$\begin{aligned}(2'''')\quad \text{Impact E + S} &= [\,Y(T) - Y(D)\,] + [\,Y(D) - Y(C)\,] \\ &< [\,Y(T) - Y(D)\,] + [\,Y^\wedge(T) - Y^\wedge(C)\,].\end{aligned}$$

As noted previously, in the first year of the analysis period (i.e., two years pre-K) outcomes for Cs and Ds are the same because neither group has as yet entered Head Start and the two groups are otherwise matched through random assignment. This circumstance allows us to substitute $Y^\wedge(D)$ for $Y^\wedge(C)$ in equation (2""), turning the right-hand side of the equation into the impact estimate defined by equation (9) and thus confirming that it is indeed an upper bound on true impact under the assumption posited. As with the lower bound in equation (7), all terms in equation (9) can be calculated from a simple 2-way experiment.

Potential Problems with This Approach

First, to develop an upper bound analogous to that in equation (9) for the kindergarten and 1st grade years of follow-up, we will have to again assume strict additivity of impacts across the two separate years of program enrollment. From equation (5), this implies that

$$(11)\quad Y(T) - Y(D) = Y(R) - Y(C), \text{ in all years,}$$

and a substitution equivalent to the replacement of $Y^\wedge(C)$ with $Y^\wedge(D)$ can again be made. [The exact derivation is fairly complex and will not be presented here.] But this makes our upper bound dependent on the same assumption as our initial point estimate. If the strict additivity assumption is true, the point estimate gives us exactly the right result and we don't need bounds (upper or lower); it the assumption is not true, we are left with nothing but a lower bound unless a new means of deriving an upper bound is found.

Second, the deferred entrants (Ds) may not come back for Head Start services a year after being turned aside by random assignment. This distorts all impact estimates that rely on $Y(D)$, though it may be possible to work around this with some added assumptions.

Finally, the calculated lower and upper bounds in equations (7) and (9) may not be well behaved (i.e., upper < lower) or of a sensible magnitude relative to the "strict additivity" point estimate.

Appendix C: Data Source for Head Start Programs in "Saturation" Communities, FACES 2000

Westat is currently collecting data for the FACES 2000 survey which, with minor modification, can supply a rich set of data on Head Start programs in "saturation" communities—communities where Head Start grantees and delegate agencies are operating at or below capacity and are unable to increase their intake flow sufficiently to provide an experimental control group. We will use data on communities of this type from FACES to test for "non-participation bias" in the randomized experiment. Several features of FACES 2000—described in this appendix—make it a good source for supplemental data of this sort.

Patterned after the original Family and Child Experiences Survey (FACES), FACES 2000 tracks thousands of Head Start children and their families over several years, collecting data on their program involvement, school readiness, family environment, and early school experiences. A total of 2,825 new Head Start participants were sampled in the fall of 2000, with 2,400 expected to remain in the program through the first wave of follow-up data collection in the spring of 2001. All ages of new Head Start entrants are represented in the sample proportionate to their prevalence in the population, including 3-year-olds starting what may be two years of Head Start enrollment and 4-year-olds beginning a single year of enrollment. (Children who had already completed a year of Head Start services were not sampled.)

The sample is spread across 43 grantees and delegate agencies and 210 centers and represents the population of Head Start participants and families nationally on a probability basis (i.e., grantees/delegate agencies and centers were sampled proportionate to size). The sample clusters in 273 classrooms, with all newly enrolled children included in the data collection for each sampled classroom. Before selection, the universe of grantees/DAs was stratified on region, urban/rural, and percentage of minorities (above or below 50 percent), as was done in the original FACES study. The FACES data also closely parallel the main experimental sample structure, coverage, and content (see text for details). Two critical differences do exist, however, one advantageous and the other a disadvantage:

- FACES 2000 includes **all** types of Head Start grantees and delegate agencies, both those that—like the experimental sites—operate at capacity in communities with enough unserved children to provide a control group and those in saturation communities that lack this characteristic; and
- The FACES 2000 data cover children entering Head Start in the fall of 2000 as opposed to the fall of 2002, and exiting in 2001 or 2002 as opposed to 2003 or 2004.

The difference in timing is only a minor drawback, since we do not expect systematic changes in Head Start program impacts over the time interval involved. The coverage of **all** of the excluded sites provides a powerful advantage.

To support the Head Start National Impact Study, FACES 2000 has been modified to collect information on grantee capacity and community saturation paralleling that to be collected in identifying saturation and under capacity sites for the main study.

Appendix D: Impact Research-related Amendment to the Head Start Act, 1998, PL 105-285

(g) NATIONAL HEAD START IMPACT STUDY.--

(1) EXPERT PANEL.--

(A) IN GENERAL.--The Secretary shall appoint an independent panel consisting of experts in program evaluation and research, education, and early childhood programs--

(i) to review, and make recommendations on, the design and plan for the research (whether conducted as a single assessment or as a series of assessments) described in paragraph (2), within 1 year after the date of enactment of the Coats Human Services Reauthorization Act of 1998;

(ii) to maintain and advise the Secretary regarding the progress of the research; and

(iii) to comment, if the panel so desires, on the interim and final research reports submitted under paragraph (7).

(B) TRAVEL EXPENSES.--The members of the panel shall not receive compensation for the performance of services for the panel, but shall be allowed travel expenses, including per diem in lieu of subsistence, at rates authorized for employees of agencies under subchapter I of chapter 57 of title 5, United States Code, while away from their homes or regular places of business in the performance of services for the panel. Notwithstanding section 1342 of title 31, United States Code, the Secretary may accept the voluntary and uncompensated services of members of the panel.

(2) GENERAL AUTHORITY: After reviewing the recommendations of the expert panel, the Secretary shall make a grant to, or enter into a contract or cooperative agreement with an organization to conduct independent research that provides a national analysis of the impact of Head Start programs. The Secretary shall ensure that the organization shall have expertise in program evaluation, and research, education, and early childhood programs.

(3) DESIGNS AND TECHNIQUES.--The Secretary shall ensure that the research uses rigorous methodological designs and techniques, (based on the recommendations of the expert panel) including longitudinal designs, control groups, nationally recognized standardized measures, and random selection and assignment, as appropriate. The Secretary may provide that the research shall be conducted as a single comprehensive assessment or as a group of coordinated assessments designed to provide, when taken together, a national analysis of the impact of Head Start programs.

(4) PROGRAMS.--The Secretary shall ensure that the study focuses primarily on Head Start programs that operate in the 50 States, the Commonwealth of Puerto Rico or the District of Columbia and that do not specifically target special populations.

(5) ANALYSIS.--The Secretary shall ensure that the organization conducting the research--

(A)(i) determines if, overall, the Head Start programs have impacts consistent with their primary goal of increasing the social competence of children, by increasing the everyday effectiveness of the children in dealing with their present environments and future responsibilities, and increasing their school readiness;

(ii) considers whether the Head Start programs--

(I) enhance the growth and development of children in cognitive, emotional, and physical health areas;

(II) strengthen families as the primary nurturers of their children; and

(III) ensure that children attain school readiness; and

(iii) examines--

(I) the impact of the Head Start programs on increasing access of children to such services as educational, health, and nutritional services, and linking children and families to needed community services; and

(II) how receipt of services described in subclause (I) enriches the lives of children and families participating in Head Start programs;

(B) examines the impact of Head Start programs on participants on the date the participants leave Head Start programs, at the end of kindergarten, and at the end of first grade (whether in public or private school), by examining a variety of factors, including educational achievement, referrals for special education or remedial course work, and absenteeism;

(C) makes use of random selection from the population of all Head Start programs described in paragraph (4) in selecting programs for inclusion in the research; and

(D) includes comparisons of individuals who participate in Head Start programs with control groups (including comparison groups) composed of--

(i) individuals who participate in other early childhood programs (such as public or private preschool programs and day care); and

(ii) individuals who do not participate in any other early childhood program; and

(6) CONSIDERATION OF SOURCES OF VARIATION.--In designing the research, the Secretary shall, to the extent practicable, consider addressing possible sources of variation in impact of Head Start programs, including variations in impact related to such factors as—

(A) Head Start program operations;

(B) Head Start program quality;

(C) the length of time a child attends a Head Start program;

(D) the age of the child on entering the Head Start program;

(E) the type of organization (such as a local educational agency or a community action agency) providing services for the Head Start program;

(F) the number of hours and days of program operation of the Head Start program (such as whether the program is a full-working-day, full calendar year program, a part-day program, or a part-year program); and

(G) other characteristics and features of the Head Start program (such as geographic location, location in an urban or a rural service area, or participant characteristics), as appropriate.

(7) REPORTS.--

(A) SUBMISSION OF INTERIM REPORTS.--The organization shall prepare and submit to the Secretary two interim reports on the research. The first interim report shall describe the design of the research, and the rationale for the design, including a description of how potential sources of variation in impact of Head Start programs have been considered in designing the research. The second interim report shall describe the status of the study and preliminary findings of the study, as appropriate.

(B) SUBMISSION OF FINAL REPORT.--The organization shall prepare and submit to the Secretary a final report containing the findings of the research.

(C) TRANSMITTAL OF REPORTS TO CONGRESS.--

(i) IN GENERAL.--The Secretary shall transmit, to the committees described in clause (ii), the first interim report by September 30, 1999, the second interim report by September 30, 2001, and the final report by September 30, 2003.

(ii) COMMITTEES.--The committees referred to in clause (i) are the Committee on Education and the Workforce of the House of Representatives and the Committee on Labor and Human Resources of the Senate.

(8) DEFINITION.--In this subsection, the term 'impact', used with respect to a Head Start program, means a difference in an outcome for a participant in a program that would not have occurred without the participation in the program.